DC-2002: METADATA FOR E-COMMUNITIES:
SUPPORTING DIVERSTY AND CONVERGENCE 2002

# Proceedings of the International Conference on Dublin Core and Metadata for e-Communities, 2002

**October 13-17, 2002 - Florence, Italy**

Organized by:

Associazione italiana biblioteche (AIB)
Biblioteca nazionale centrale Firenze (BNCF)
Dublin Core Metadata Initiative (DCMI)
European University Institute (IUE)
Istituto e Museo di Storia della Scienza (IMSS)
Regione Toscana
Università degli Studi di Firenze

**Firenze University Press**

Proceedings of the International Conference on Dublin Core and Metadata for e-Communities, 2002

FUP

DC-2002: METADATA FOR E-COMMUNITIES:
SUPPORTING DIVERSITY AND CONVERGENCE 2002

# Proceedings of the International Conference on Dublin Core and Metadata for e-Communities, 2002

**October 13-17, 2002 - Florence, Italy**

Organized by:

Associazione Italiana Biblioteche (AIB)
Biblioteca Nazionale Centrale di Firenze (BNCF)
Dublin Core Metadata Initiative (DCMI)
European University Institute (IUE)
Istituto e Museo di Storia della Scienza (IMSS)
Regione Toscana
Università degli Studi di Firenze

**Firenze University Press**

DC-2002 : metadata for e-communities : supporting diversity and convergence 2002 : proceedings of the International Conference on Dublin Core and Metadata for E-communities, 2002, October 13-17, 2002, Florence, Italy / organized by Associazione italiana biblioteche (AIB) … [at al.]. – Firenze : Firenze University Press, 2002.
http://epress.unifi.it/

ISBN 88-8453-043-1
025.344 (ed. 20)
Dublin Core – Congresses
Cataloging of computer network resources - Congresses

*Print on demand is available*



Firenze University Press
Borgo Albizi, 28
50122 Firenze, Italy
http://epress.unifi.it/

*Printed in Italy*

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002*


# DC-2002: Metadata for e-Communities: Supporting Diversity and Convergence

DC-2002 marks the tenth in the ongoing series of International Dublin Core Workshops, and the second that includes a full program of tutorials and peer-reviewed conference papers. Interest in Dublin Core metadata has grown from a small collection of pioneering projects to adoption by governments and international organizations worldwide. The scope of interest and investment in metadata is broader still, however, and finding coherence in this rapidly growing area is challenging.

The theme of DC-2002 — supporting diversity and convergence — acknowledges these changes. The greatest challenge of the current phase of metadata development is bringing together the diversity of local conventions, domain specific requirements, and different encoding conventions such that cross-domain interoperability can be achieved.

DC-2002 brings together practitioners from libraries, museums, education, government, commerce, supra-governmental organizations, and others in a venue for discourse, education, exploration, and development of metadata themes on the Web and elsewhere. Contributions come from practitioners focused on the needs of constituents on the one hand and theoreticians exploring the 'Semantic Web' on the other. Adopters are faced with decisions about XML, RDF, Schema languages, ontologies, Open Archive Initiative protocols, and more. The richness of this diverse field of endeavor is both exciting and daunting.

It is fitting that DC-2002 takes place in Florence, in the shadow of Il Duomo. Brunelleschi's dome, still acclaimed as one of the world's great buildings, offers us lessons for our own efforts. Francesco Talenti's 1366 design for Il Duomo called for a dome that surpassed the limits of architectural practice of the day. It was more than 50 years before Brunelleschi's design to complete the building was approved and years more before it could be executed.

Our metadata systems may never achieve the grace and beauty (let alone the persistence!) of this great hallmark of Renaissance confidence, but we share with these visionaries acceptance of uncertainty and the confidence to act in spite of it. The authors of the papers in these Proceedings, and the conferees that engage them, will play a part in resolving these uncertainties and strengthening the foundations of resource discovery.

Stuart Weibel
Director, Dublin Core Metadata Initiative

# Message from the Organising Committee

DC-2002, the International Conference on Dublin Core and Metadata Applications 2002 is taking place in Florence. In late 2001 Dublin Core Metadata Initiative (DCMI) asked a number of us if we saw a possibility to host this annual event . One of the ideas behind the choice of location was the intention to highlight the metadata concept for a potential public in Southern Europe. We accepted this challenge and have worked to make this year's Conference an occasion where experts, practitioners as well as learners can profitably come together and take away new ideas and insights. With the array of keynotes, papers, posters, workshops, tutorials, special topic sessions and birds-of-a-feather meetings on offer every participant will find plenty of interest.

Later this same week the annual Italian Library Conference is taking place in Rome. We have profited from the presence of this large gathering of Italian librarians to organize an extra Dublin Core introductory tutorial, in Rome and in the Italian language. Interest has been overwhelming.

In addition to the conference programme itself the organisers have tried to offer the participants an experience that will make their brief stay in Florence agreeable. The Conference venue, the Oltrarno Meeting Center, a former convent by the name of "Convitto della Calza", is a beautiful and historical place. Other historical buildings in the town of Florence will host some of the social events during the week.

Although the DC-2002 Conference is largely self-financed we need to thank a number of organisations for their material, financial, and staff support. These are listed in the next pages.

Finally, many individuals have given of their time and effort in order to create the conditions of success for this Conference. We owe them our thanks. Among them a few certainly deserve special mention because of their constant and unstinting involvement: Massimo Rolle, Susanna Peruginelli and Giovanni Bergamin.

On behalf of the local Organising Committee I wish all participants a most successful conference.

Michiel Tegelaars
Coordinator of the local Organising Committee

THE DC-2002 CONFERENCE IN FLORENCE WAS ORGANIZED UNDER THE AUSPICES OF THE DUBLIN CORE METDATA INITIATIVE BY THE FOLLOWING PEOPLE:

**Programme Committee members**

Jun ADACHI
Isidro AGUILLO
Ann APPS
Ray ATARASHI
Tom BAKER
Miroslav BARTOŠEK
Olga BARYSHEVA
Nicola BENVENUTI
José BORBINHA
Nancy BRODIE
Pino BUIZZA
Debbie CAMPBELL
Hsueh-hua CHEN
Geneviève CLAVEL
Gianfranco CRUPI
Roberto DALLARI
Michael DAY
Antonella DE ROBBIO
Makx DEKKERS
Sandra DI MAJO
Asuman DOGAC
Antonella FARSETTI
Elisabeth FREYRE
Janifer GATENBY
Simona GATTA
Dino GIULI
Claudio GNOLI
Jane GREENBERG
Mariella GUERCIO
Mauro GUERRINI (Co-chair)
Juha HAKALA
Jytte HANSEN
Preben HANSEN
Liv HOLM
Jane HUNTER
Stefan JENSEN
Pete JOHNSTON
Noriko KANDO
Stephen KATZ
Johannes KEIZER
Sung-Hyuk KIM
László KOVÁKS
John KUNZE
Cristina MAGLIANO
Stefania MANZI
Lynn MARKO
Enrico MARTELLINI
Liddy NEVILE
Ayling ONG
Federica PARADISI
Dina PASQUALETTI
Valdo PASQUI
Maria PEPE
Ginevra PERUGINELLI
Susanna PERUGINELLI
Maria Chiara PETTENATI
Siripan PRADITTA
Mario RAGONA
Oh SAM
Antonio SCOLARI
Carla SIMONE
Aida SLAVIC
Shigeo SUGIMOTO
Anna Maria TAMMARO
Federico VALACCHI
Stefano VITALI
Stuart WEIBEL (Co-chair)
Andrew WILSON
Eric ZIMMERMAN

**Publicity Committee**
Giovanni BERGAMIN (National Library, Florence)
Dina PASQUALETTI (National Library, Florence)

**Publication Committee**
Lucia BERTINI (Firenze University Press)
Patrizia COTONESCHI (Firenze University Press)
Antonella FARSETTI (Firenze University Press)

**Tutorial Committee**
Makx DEKKER (DCMI)

**Local Organizing Committee**
Catherine DEELEY, Secretary
Massimo ROLLE (Regione Toscana)
Michiel TEGELAARS, Co-ordinator (European University Institute)

THE DC-2002 CONFERENCE IN FLORENCE RECEIVED MATERIAL SUPPORT FROM THE FOLLOWING ORGANIZATIONS:

Associazione Italiana Biblioteche
Biblioteca Nazionale Centrale di Firenze
Fondazione Cassa di Risparmio di Firenze
Comune di Firenze
Dublin Core Metadata Initiative
European University Institute
Florence University Press
Istituto e Museo della Storia della Scienza
Regione Toscana
Università degli Studi di Firenze

# Contents

## Paper Session 4



## Posters

Paper Session 1

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 7-17*


# District Architecture for Networked Editions: Technical Model and Metadata

Antonella Farsetti
Firenze University Press (FUP) - Università di Firenze
e-mail: antonella.farsetti@unifi.it

Valdo Pasqui
Centro Servizi Informatici dell'Ateneo Fiorentino (CSIAF) - Università di Firenze
e-mail: valdo.pasqui@unifi.it

## Abstract

*District Architecture for Networked Editions (DAFNE) is a research project funded by the Italian Ministry of Education, University and Research aiming to develop a prototype of the national infrastructure for electronic publishing in Italy. The project's initial target concerns the scientific and scholarly production in the human and social sciences. The organizational, legal, technical and business aspects of the entire digital publishing pipeline have been analysed. DAFNE system will support the request-offer chain by promoting the integration between the digital library and the electronic publishing districts. In this paper we present the main results of the project's first year of activity. First a quick outlook about the actors, objects and services is presented. Then the functional model is examined bringing out the distinction between information content and digital objects. Afterwards the technical model is described. The system has a distributed architecture, which includes three categories of subsystems: Data Providers (i.e. the publishers), Service Providers and External Services. Data and Service Providers interact according to the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Finally DAFNE metadata is discussed. Metadata permeates the whole publishing chain and DAFNE metadata set is based on already defined domain-specific metadata vocabularies. Dublin Core Metadata Initiative (DCMI) and Publishing Requirements for Industry Standard Metadata (PRISM) are the main reference standards. Open Digital Rights Language (ODRL) and Open Archival Information System (OAIS) are the two other relevant models which complete DAFNE metadata specification.*

**Keywords:** *electronic publishing, data and service provider, descriptive metadata, rights management, digital objects, OAI-PMH.*

## 1. Introduction

During the last few years the publishing industry has been greatly affected by Internet and its new technologies. All the main publishers propose Web portals to organize their catalogues and services, including the possibility to read on-line or to download publications in electronic format. Many journals have a digital on-line version or they are published only in electronic format (i.e. born digital). The standardization of metadata schemas, in particular thanks to Dublin Core Metadata Initiative (DCMI), and the wide acceptance of the Protocol for Metadata Harvesting (PMH) promoted by the Open Archives Initiative (OAI), are pushing the creation of Internet-based services in the publishing market. The consolidation of XML standards and tools is leading to a more structured organization of the contents, independent of proprietary formats such as Adobe PDF and MS Word. Furthermore, the quick diffusion of e-commerce platforms is promoting more flexible trading models in the publishing market. Finally, in the academic and research communities relevant initiatives aim to promote the "free" diffusion of the scholarly knowledge by the creation of e-prints services, mainly based on institutional and thematic archives.

In this framework, the Italian publishing state-of'-the-art is still in its infancy. Moreover, some sectors such as the economy and the social sciences are self-consuming knowledge-generation contexts. This is a niche market whose authors and consumers are the same actors. In particular the distribution of scientific journals has its main focus on subscribers, special clients and specialized bookshops.

District Architecture for Networked Edition (DAFNE) is a research project funded by the Italian Ministry of Education, University and Research (MIUR) which aims to develop a prototype of the national infrastructure for electronic publishing in

Italy. The project takes into account the organizational, legal, business and technical aspects in order to define a set of services and tools for supporting the subjects involved in the digital publishing pipeline. The main goals of DAFNE are: i) to improve the efficiency and efficacy of the request-offer chain; and ii) to promote the integration between the digital library and the electronic publishing districts.

Firstly the project aims to analyse the scholarly and scientific production in the human and social sciences.

The project's participants are five relevant Italian companies: Ariadne (ICT), Bassilichi (e-commerce), Editrice Bibliografica (publisher), Casalini (bookseller) and the consortium "Parco Scientifico e Tecnologico Galileo" (ICT). The scientific committee includes three main Italian universities (Bologna, Firenze and Padova), the "Istituto di Teoria e Tecnica dell'Informazione Giuridica" of Consiglio Nazionale delle Ricerche (CNR), the Biblioteca del Mulino and the Biblioteca Nazionale Centrale di Firenze (BNCF). The University of Florence is participating in the project with the Firenze University Press responsible for metadata specification and the CSIAF responsible for the technical design.

DAFNE started in October 2001 and will last for three years. Three reference models have been defined in the first year: legal, organizational and technical. Relevant metadata standards have been analysed and a draft proposal for DAFNE reference metadata has been elaborated. Both the metadata set and the business model will be finalized by the end of the current year.

## 2. Actors, objects and services

### 2.1 Actors

In the electronic publishing pipeline one end includes a set of processes such as content acquisition and generation, digitisation and layout editing. At the other end of the pipeline there are the business transactions with the consumers (i.e. end-users). The DAFNE organization model has stressed the need for an Internet-based system to support the consultation (metadata and contents) and the on-line sale of electronic publications. DAFNE aims to provide tools and services to create a bridge between the publishers and the end-users in order to improve their relations: consumers can evaluate the contents of works more in depth and publishers can take into account the preferences of their clients. In this context the main actors, including both human subjects and external systems, are:

- authors, who generate intellectual contents
- publishers, including editorial-staff members
- end-users, the consumers (teachers, researchers, students, private companies members, professional experts, etc.) who look for publications about their scientific, research or personal interests
- brokers, service providers, public and academic libraries which support value-added services such as metadata, abstracts and TOCS creation, metadata and full-text indexing, document access
- national libraries that manage legal deposit services
- systems for the assignment and resolution of document unique identifiers (e.g. ISBN, ISSN, NBN, DOI)
- e-commerce services, such as banks and credit-cards networks, to support on-line payments and to manage the accounting of intellectual propriety rights
- Public Key Infrastructure (PKI) services, including Certification Authorities, to support cryptography, strong authentication and digital certificates
- delivery services, such as national postal systems and private couriers

### 2.2 Objects

In the academic context a wide range of digital objects can be organized in electronic archives or directly published on-line. Three reference axes have been explored for all documents: types, aggregation level, and formats.

*Document types*. On the basis of the Firenze University Press experience the following kinds of documents have to be considered:

– monographs and conference proceedings
– electronic journals
– pre-prints and post-prints of journal papers
– technical and research reports
– lecture notes, teaching notes and learning material
– multimedia publications, including audio-visual recordings
– collections of digitised documents
– theses and dissertations
– bibliographic and review collections about specialized thematic topics

*Aggregation level*. In the previous list many documents are basically monographic. They are standalone publications or can be collected in series depending on specific publishing policies. Multimedia publications can be organized, as monographic issues and their manifestation can be a file to be downloaded or a full hypermedia document to be navigated and read/played/seen on-line. Usually, bibliographic and review collections are managed by databases and a set of procedures which support a search and retrieve Web interfaces. On-line journals keep the same aggregation levels of their related paper version, that is the volume/number hierarchy. So the end-user navigates by year, then by single issue that contains the table of contents pointing to articles. These papers are described by a basic set of

metadata, which contains the links to the full-text version, usually restricted by access control mechanisms such as user registration or IP address authentication. The contents of born-digital journals (e.g. D-Lib Magazine, Ariadne, IEEE Distributed Systems ONLINE) are more structured and their graphic layout is generally richer. In any case descriptive metadata must be associated to documents to represent their relationships. In particular, keywords and subject elements make up "virtual" aggregations by thematic areas based on selected subject and classification schemas. Relation elements must be used to express at least hierarchical relations like "is-part-of" and "has-part".

*Formats*. Basically DAFNE will support the commonly diffused document formats: HTML, Adobe PDF, Microsoft Word, RTF and TeX (used in mathematics and physics). The lack of consolidated e-books standards has suggested postponing these new kinds of publications for future investigation. In the advanced model the DAFNE production phase is two fold, both based on XML standards. The former deals with the generation of the information contents according to pre-defined XML DTD and schemas. The latter includes the layout formatting and the graphical rendering, even in different formats such as HTML and PDF, by means of a transformation process based on the Extensible Stylesheet Language (XLS) and XML Transformations (XLST). The first phase has the focus on the content and its structure, independent of proprietary formats. Descriptive metadata (e.g. Dublin Core elements) can be included to be automatically extracted in the next processing steps. The second phase concerns the possibility of generating different manifestations of the same intellectual work. In this step preservation and technical metadata should be added to properly manage the generated bit stream.

### 2.3 Services

Publishers like Addison Wesley, Elsevier or Springer are typically oriented to business support by offering virtual on-line shops to their consumers. Sometimes these Web sites are integrated with e-commerce sub-systems based on the shopping basket model. In other cases, consumers are redirected to booksellers and distributors sites such as Amazon.com and Fatbrain to buy on-line. The recent development of portals technology started a new generation of sites that support user profiling and monitor their activities. Thus the end-user operating environment can be dynamically tailored on the basis of her/his preferences and habits. Publishers are pushing consumers to join their portals as one-stop shop points for discovering, retrieving and accessing all the information they search for in the net.

DAFNE's focus is on the generation and diffusion of electronic publications in the scientific and academic context. For several years free consultation services have been exploited in this area such as ArXiv, NCSTRL and CogPrints. These e-print systems are based on the author/institution self-archiving model with dedicated Web interfaces for the direct submission of documents and the creation of metadata by registered authors. As regards papers the full-text visualization and print is free for the end-users.

Starting from this realm, since October 1999 the Open Archives Initiative (Van de Sompel & Lagoze 2001) has defined a general framework for open archives metadata interoperation based on the concepts of Data and Service Providers and the Protocol for Metadata Harvesting (PMH) now available in version 2 (OAI). The Electronic and Computer Science Department of the University of Southampton has implemented Eprints, a software for the creation and management of e-prints archives based on the self-archiving approach. Eprints version 2 supports the last PMH version and is free for download, compliant to GNU guidelines.

Experts like Steven Harnad (Harnad 1999) and Paul Ginsparg (Ginsparg 2001) have undertaken a true mission for promoting e-prints open archives in the scholarly publishing context as the solution for the rapid and free diffusion of scientific knowledge. Recently the Scholarly Publishing & Academic Resources Coalition (SPARC) has issued a paper to stimulate the discussion about scholarly communication and institutional repositories. The report contains two relevant positions: a) "institutional repositories can provide an immediate complement to the existing scholarly publishing model, while stimulating the emergence of a new disaggregated publishing model that will evolve over time"; and b) "Institutional repositories represent the logical convergence of faculty-driven self-archiving initiatives, library dissatisfaction with the monopolistic effects of the traditional and still-pervasive journal publishing system, and availability of digital networks and publishing technologies." (SPARC 2002, p. 29).

DAFNE, taking into account these two models, publishers "business shop" and academic free "knowledge diffusion", has defined a flexible framework to let both live together for pay publications and free e-prints archives (e.g. pre-prints, post-prints, lectures and learning material). In particular this approach is suitable for those University Presses which publish several kinds of documents and have two main categories of consumers: "internal" users (enrolled students, teachers, researchers) and "external" buyers (professional specialists, private companies, common citizens). The analysis phase has pointed out the following set of services to be supported by the system:

- document submission in digital format
- related metadata creation and upgrading
- peer-reviewing
- document digital signature, time stamping and encryption

– full-text indexing
– metadata indexing, search and retrieval
– authors registration and authentication
– end-users registration, profiling and authentication
– documents (e.g. full-text) access according to different kinds of policies (e.g. pay-per-view, print-on-demand, download, subscription)
– alerting, news, interest group subscription
– on-line payment (e.g. by credit cards)
– copyright management and accounting of royalties related to intellectual proprietary rights

## 3. The functional model

One of the main prerequisites of the project was to define a reference framework to manage several types of documents, to allow different organization/ aggregation models and to exploit flexible archiving systems. Dealing with the academic and scientific publishing context the model must support:

– a high level of autonomy for the authors who generate the intellectual contents
– local, national and international interoperability of document collections by thematic areas
– the integration with other information resources (primary and secondary) to support end-users uniform and simple access
– co-existence with business publishing

These goals have implied a clear distinction between two concepts: information (or intellectual) content and digital objects. The former is an abstract concept; the work generated by the intellectual activity of its authors, which is the core of the academic and scientific production, independent of formats, organizations and access modalities. The latter is the concrete representation of the information content within the system. According to this logical view, a digital object becomes an extension of the information content and includes the following components:

a) metatada which describes all the features of a digital object; at least four categories of metadata must be included: descriptive, representation and technical data, copyright and intellectual properties data, end-users' access rights;
b) the physical representation of the information content, formed by one or more bit-streams (i.e. texts, sounds, images, video sequences);
c) a persistent and unique identifier assigned when the digital object is created in the system.

In DAFNE the publishing pipeline has been partitioned in a sequence of relevant logical phases (Pasqui 2001). The submission of a new information content is the starting point A dedicated system module supports the direct immission of documents by pre-registered authors who also create a basic set of descriptive metadata. This activity can be executed by the editorial-staff (e.g. when a document is sent by e-mail). In any case, they perform formatting, graphical and layout restyling by using off-the-shelf authoring tools. Moreover they are in charge of the revision and full creation of descriptive metadada that feed the on-line catalogue to be searched/ browsed by end-users. At the end of this phase the information content becomes a digital object stored in a dedicated area of the publisher's repository, accessible only to the authors, the editorial staff and, if necessary, to the reviewers.

The peer-review is a well-known activity in the scientific publishing context. To manage and track all the interactions with the reviewers a "by hand" process can be used, using e-mails to exchange documents and review comments. Otherwise a dedicated system, such as Manuscript Central (ScholarOne), Xpress Track (XpressTrack), EdiKit (Berkeley Electronic Press) can be used. DAFNE's first prototype follows the first approach and the integration with an off-the-shelf tracking system has been planned in the advanced version.

When the information content is ready to be issued for publication the editorial-staff has to conclude the processing by performing several other operations. First, all the metadata related to the new publication has to be added: descriptive (including subject and relation entities), technical, copyright, end-user rights (including access modalities and related costs). Second, a persistent and unique identifier has to be assigned to the publication based on the Uniform Resource Name (URN) standard (IETF RFC 2141). URNs are persistent, location-independent, resource identifiers, a subset of the more general category of resource identifiers known as Uniform Resource Identifiers (URI) (IETF RFC 2396). In the publishing area exists a de-facto standard, the Digital Object Identifier (DOI) system, promoted by an international coalition of commercial publishers. Being a research project, DAFNE is investigating the possibility to develop a light identification system based on Light Directory Access Protocol (LDAP) technology. Third, a digital signature and a digital timestamp (based on PKI technology) can be generated to assure document authenticity and not repudiation by authors. In this phase the staff can submit the publication, including its digital signature and a subset of metadata, to the national service responsible for the legal deposit of electronic publications, which in Italy is the Biblioteca Nazionale Centrale di Firenze. Finally, the digital object enters the persistent storage area of the publisher's repository. This means that its related metadata becomes available for consultation and export. Now the publication is ready for access on the basis of the rights permission and control access mechanisms defined before.

In DAFNE publisher's catalogue consultation (i.e. end-users search and discovery) and other services such as alerting, news diffusion and even user regis-

tration and access control can be supported by the publisher directly or by a third party service provider, acting as a broker between the consumers and the publisher. In the first case the catalogue and the other services are integrated in the publisher repository, in the latter all the services have been delegated to a service provider and the publisher simply acts as a data provider which exports its metadata and objects.

Digital object access is the functional component, which deals with access control policies and business logic. Access control is based on the matching of two classes of properties: i) the access rights modalities associated to each digital object which are described by dedicated metadata; ii) the attributes which characterize each end-user (e.g. type, institution enrolment, subscriptions and registrations). Unless a publication is free, some kind of user identification and authentication should be implemented (e.g. IP based for institutional users and user/password for generic consumers). Business logic deals with all the aspects related to payments and royalties accounting. To this end the on-line catalogue Web interface must be integrated with an e-commerce subsystem which supports on-line payment transactions. DAFNE will exploit a synchronous approach that immediately notifies the seller's server about the successful completion of on-line payment. This allows the implementation of the pay-per-view access model without user subscription or pre-registration. Credit cards and mobile telephone payment models will be experimented. As far as royalties accounting is concerned DAFNE functional model asks for specific metadata to describe the fee (e.g. a fixed amount or a percentage of the cost) and to identify the subjects (persons, institutions or companies) entitled to that fee. These subjects must have a persistent and unique identifier. To avoid the maintenance of local repositories, a central (national) service for the registration of the subjects (e.g. authors) involved in copyright management should be implemented. Publishers and service providers use these metadata to compute the royalties resulting from end-users access to digital objects. The cumulated amounts will be periodically transferred to the central system to generate the payment transactions to be credited to the entitled subjects by a bank.

## 4. The architecture

The overall system architecture is distributed according to a three level logic depicted in Fig. 1 as a sequence of concentric rectangles. Data Providers are the core level, which includes basic services such as digital objects repositories. In DAFNE a publisher is a data provider. At the intermediate level there are the Service Providers, which manage value added services (e.g. resource cross-references). Moreover, they can support basic services delegated by publishers. The Support Services layer includes autonomous systems that provide specific services, shared by the whole publishing community on the basis of national and inter-institutional agreements.

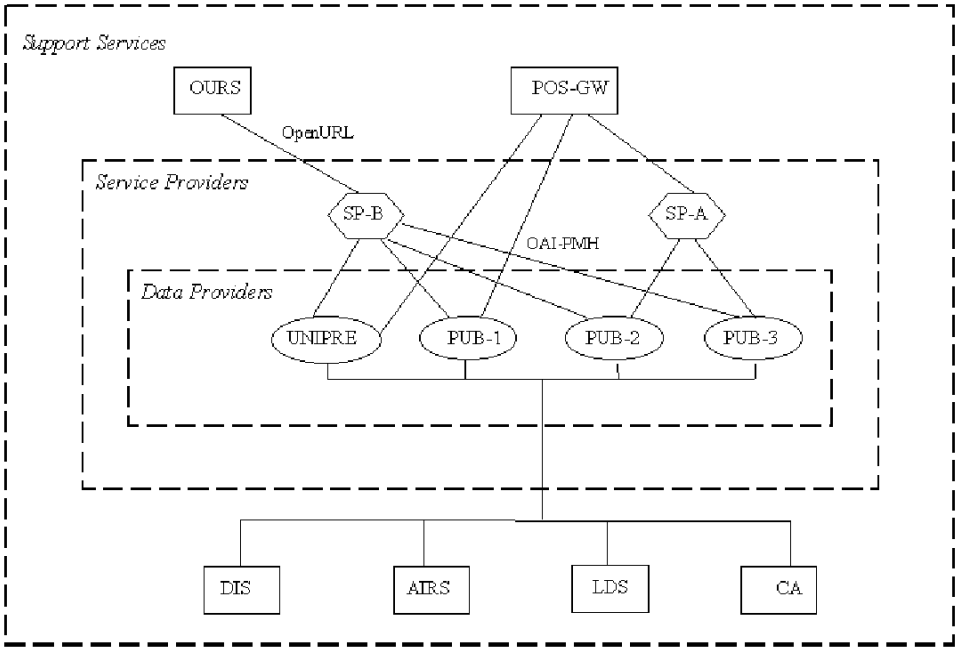



**Figure 1. Three-level architecture and a possible deployment scenery**

The design of Data and Service Providers subsystems is fully modular and their implementation should be component based to support the maximum flexibility and scalability.

Data Provider components are:

- *Storage* [+], the repository for the storage of documents and metadata.
- *Submission* [+], a Web interface for documents submission.
- *Authors Registration&Authentication* [+], the module responsible for authors registration and authentication.
- *PMH* [+], the interface that supports metadata harvesting from the repository according to PMH v.2 specification.
- *Document Identification Interface* [+], the module that supports the interaction with the service responsible for the assignment and resolution of persistent unique identifiers to digital objects.
- *Peer-review*, the tracking system to manage the reviewing activity.
- *End-Users Registration&Atuthentication*, the module responsible for authors registration and authentication.
- *End-Users Search Interface*, the Web interface to the catalogue which provides both simple and advanced search, main index browses, navigation by series and other documents aggregations.
- *End-Users Authorization*, the module that checks end-users access to digital objects.
- *E-commerce module*, the component for on-line payments management, interfaced to the POS Gateway subsystem.
- *Copyright module*, the component which implements rights management fees computation and accounting.

Only the basic components, marked with [+], must be deployed by a publisher (i.e. a data provider). The other modules can be omitted if their functions are delegated to an external service provider. Each Data Provider must implement the Protocol for Metadata Harvesting. This module supports any interaction with the external world, in particular with Service Providers, to export the metadata related to the digital objects hosted in the repository.

Fig. 2 is the schematic representation of a full-component Data Provider subsystem.

Service Provider components are:

- *Storage*, to host the metadata harvested from data providers' repositories.
- *Cross-reference linking*, to extend the items with links to other related resources (e.g. OPACs, bibliographic databases, abstracting services) based on the OpenURL standard (Van de Sompel & Beit-Arie 2001).
- *PMH*, the interface to harvest metadata from Data Providers repositories
- *End-Users Registration&Authentication*, the module responsible for authors registration and authentication.
- *End-Users Search Interface*, the Web interface to the catalogue that provides simple and advanced search functions, main index browses, navigation

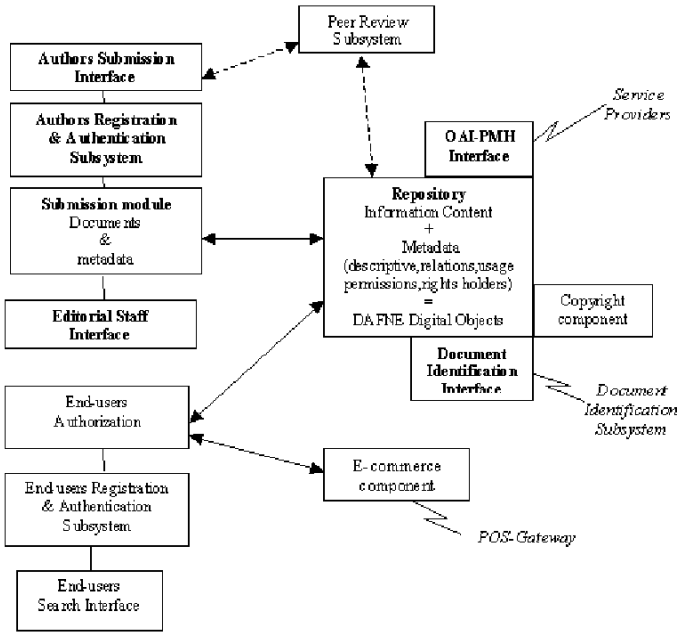



**Figure 2. Full Data Provider components**

by series and other document aggregations.
- *End-Users Authorization*, the module that checks end-users access to digital objects.
- *E-commerce module*, the component for on-line payments management, interfaced to the POS Gateway subsystem.
- *Copyright module*, the component that implements rights management, fees computation and accounting.

A Service Provider can support many publishers. For instance many independent catalogues can be implemented. In this case orders relating to items from different publishers catalogues can be submitted. The modules which interface the digital object identifier resolution subsystem, the POS Gateway and the copyright accounting system are unique.

The Support Services are:

- *Document Identification Subsystem,* which supports the assignment and resolution of persistent-unique identifiers for digital objects, conformant to URN syntax.
- *POS-Gateway* to credit card circuits, a system to manage on-line payment transactions, usually hosted by a bank.
- *OpenURL Resolution Service,* a centralized subsystem which implements the knowledge base and the resolution mechanisms to provide proper (i.e. context sensitive) links related to a catalogue item activated by an end-user.
- *Authors Identification&Registration Subsystems,* which support the registration and unique identification of the subjects, entitled to copyright royalties. The related database contains personal data, including the bank coordinates for the automatic generation of credit transactions.
- *Legal Deposit Service*, a centralized subsystem responsible for the legal deposit of electronic documents at the national level.
- *Certification Authority,* which issues, revokes and renews digital certificates to support digital signatures and timestamps.

In Fig. 1 a possible scenery showing the main interactions is depicted:

## 5. Metadata

As widely defined in literature, metadata is data about data and "includes information about the context of data and the content of data and the control of or over data" (Pasquinelli 1997). Metadata permeates the entire life cycle of digital publications. As the functional and technological models show, DAFNE is totally plunged into metadata, which supports the retrieval, management, long term archival, use and reuse of its digital objects.

One of the basic prerequisites of the project was to

| | |
|---|---|
| UNIPRE | Is a University Press, which publishes several kinds of documents included pre-prints for free consultation and learning material, restricted to campus network. Registered teachers and researchers can self-archive their papers. The other publications are subject to payment and subscription policies. |
| | UNIPRE implements all the modules defined for a Data Provider except the review subsystem. |
| PUB-1 | Is a commercial publisher whose catalogue contains digital journals and monographs. PUB-1 exploits a peer-review subsystem. IP address check is the control access policy for journals, whereas on-line payment is requested for subscriptions and pay-per-print. |
| | PUB-1 implements all the modules defined for a Data Provider. |
| | PUB-2 and PUB-3 Are two commercial publishers which support the direct submission of papers by authors. They have delegated their catalogue interface and on-line payment to a service provider. |
| PUB-2, PUB-3 implement only the basic Data Provider components. | |
| SP-A | Is a Service Provider which implements consultation, access control and on-line payment for PUB-2 and PUB-3. |
| | Metadata is captured from the publishers by PMH. |
| SP-B | Under the conditions of a special agreement, this Service Provider integrates the collections from UNIPRE, PUB-1, PUB-2 and PUB-3 by using PMH. Moreover, SP-B adds to the items an OpenURL link for cross-referencing with other resources, resolved by a dedicated subsystem (OURS). |
| OURS | The OpenURL Resolution Service. |
| DIS | The Document Identification Subsystem. |
| POS-GW | The POS-Gateway. |
| AIRS | The Authors Identification&Registration Subsystem. |
| LDS | The Legal Deposit Service. |
| CA | A Certification Authority. |

identify a metadata set without creating anything new, but just by using metadata vocabularies already defined in related domains. In order to define an appropriate metadata profile, research and analysis of existing metadata sets have been conducted. As far as interoperability issues are concerned, this approach promotes the integration between the digital library and the electronic publishing districts. Moreover, we must not forget the great amount of resources, which have been invested in the definition of common metadata, sets in the digital objects realm. DCMI is part of this effort. Therefore it is from Dublin Core (DC) that DAFNE metadata analysis has begun to go on with other specifications.

The Open Archival Information System (OAIS 1999) has guided the design of the DAFNE information model, helping to select the metadata sets relevant in the digital publishing pipeline.

OAIS defines a conceptual framework for generic archival systems and an information model that provides a systematic view of metadata gathered in several categories. This model is "a high-level description of the types of information generated by and managed within the functional components of a complete archiving system" (OCLC/RLG 2001). The archive information package includes four types of information objects: Content Information, Preservation Description Information, Packaging Information and Descriptive Information. Content Information consists of the digital object and its associated representation information, that is the technical metadata, which supports the correct rendering and interpretation of the associated digital object (i.e. bit stream). The Preservation Description Information includes four types of metadata: Provenance (about the origin and preservation), Reference (includes the identifiers associated to a digital object), Fixity (data related to the authentication mechanisms) and Context Information (metadata about relations with other objects). Descriptive Information supports search and retrieve of the archive contents.

The first phase of DAFNE focuses on descriptive and digital rights management metadata. Technical, structural and long term preservation metadata will be analysed in depth in a further phase of the project. Concerning descriptive metadata, Publishing Requirements for Industry Standard Metadata (PRISM) is the reference metadata set in DAFNE, in particular for the design of data providers' repositories. PRISM is a specification promoted by a group of publishers and vendors, joined under the IDEAlliance initiative, which defines a set of elements useful for "interoperable content description, interchange and reuse in both traditional and electronic publishing contexts" (PRISM 2002, p. 1). This specification recommends the use of existing standards such as Dublin Core, XML, RDF and various ISO standards for location, languages and date/time format. PRISM descriptions are expressed as standalone XML documents or PRISM metadata can be embedded inline within the document. The XML representation is totally compatible with OAI-PMH and its capability to transport multiple metadata sets.

PRISM elements are gathered by six functional groups: General Purpose Descriptive Elements, Provenance, Timestamps, Subject Description, Rights and Permissions, Resource Relationships. The last group is proposed in place of "dc:relation" to provide more peculiar definitions about relations among the resources. A series of controlled vocabularies (PCV) enrich PRISM elements: Rights and Usage (prl:usage) to specify resource usages, Resource Type to define the style of presentation in the resource's content (dc:type), Resource Category to specify the intellectual genre (prism:category). PCVs support a further functional description for basic use of documents. The list of terms in Resource Type Vocabulary and Resource Category vocabularies derive from third parties thesauri such as Arts and Architecture Thesaurus (AAT), WORLDNET and NewsML (PRISM 2002, p. 52-55).

In the digital environment the creation of intellectual content is tightly linked to its management and commercial use and reuse, where "commercial" is used in its broadest sense. Commerce for DAFNE includes trade with the consumers and "cultural transactions" with public libraries or other not for profit institutions (e.g. universities, schools, etc.). In this realm, any intellectual content that transforms into a digital object can be related to many actors claiming rights on it. Indecs project (<indecs>), which deals with identifiers and metadata supporting intellectual property rights, asserts that "while an apple bought at a market stall is a single physical entity owned entirely by one person, a single digital audiovisual creation may contain hundreds or even thousands of separate pieces of intellectual property." (Rust & Bide 2000, p. 4). This clearly identifies how complex digital objects and their Intellectual Property Rights (IPR) management are.

As remarked by Iannella (Iannella 2001), Digital Rights Management (DRM) is the "digital management of the rights", be they the rights in a physical manifestation (e.g. a book) or be they the rights in a digital manifestation (e.g. an e-book) of a work. The first generation of DRM systems aimed to prevent unauthorized copies of digital objects by security controls and encryption techniques. Second generation DRMs include the description, identification, trading, monitoring and tracking of rights and the relations with the entitled subjects.

Waiting for the consolidation of one of the emerging metadata standards for DRM, such as the Xtensible rights Markup Language (XrML) and the Open Digital Rights Language (ODRL), PRISM adopted a pragmatic approach. The Rights and Permission functional group specifies "a small set of elements that would encode the most common rights information to serve as an interim measure for inter-

operable exchange of rights information" (PRISM 2002, p. 14). This set is too limited to support the basic level of DRM needed in DAFNE, so other relevant models have been investigated. Starting with <indecs> project, DAFNE has selected ODRL as the reference model for rights management metadata. Basically ODRL is a language and a vocabulary to express terms and conditions over assets. An asset is any digital or physical intellectual content, over which the author or other parties can claim rights. "ODRL complements existing analogue rights management standards by providing digital equivalent and supports an expandable range of new services that can be afforded by the digital nature of the assets in the Web environment." (Iannella 2001a, p. 1). ODRL defines a model and a set of semantics for managing the rights holders and the permissible usages of asset manifestations. Thanks to ODRL any digital asset can have its digital rights management information linked to it.

According to ODRL specified semantics, some of its data dictionary elements can integrate PRISM helping to define DAFNE metadata reference set, mainly:

- Usage Permission elements (*display*, *print*, *play*, *execute*)
- Requirement elements (*payment* which contains amount and currency and tax percent and code, *prepay*, *postpay*, *peruse*)
- Rights Holder elements (*percentage*, *fixedamount*)

Permissions are linked to parties and assets through an *agreement* element. Requirements are associated to permissions. Rights Holder elements are included within party elements, the subjects entitled to royalties. Both assets and parties must have a unique identifier, which is expressed by the *context* element and its sub-elements.

DAFNE Metadata Profile is on the way to be finalized. The Appendix outlines the core elements set under specification, derived from DC, PRISM and ODRL. The listed identifiers follow the XML Namespaces syntax where "dc" and "prism" are the namespaces that include Dublin Core and PRISM elements, whereas "dafne" should be the name of a new metadata vocabulary to be defined. DAFNE repository implementation (i.e. the Storage component) will put in relation Content Information, Metadata, Parties claiming rights on them, Permissions, Usage Constraints and Requirements. XML syntax is used to define simple and complex elements relations.

## 6. Conclusions

DAFNE is a research project that concerns scientific and scholarly production in the human and social sciences. Aiming to define the prototype of the national infrastructure for electronic publishing in Italy a full analysis of the publishing pipeline was performed. This has led to the definition of the organizational, legal, technical and business models. This paper has outlined the functional model, the reference architecture and the core set of metadata. DAFNE deals with the academic realms within which, its functional model tries to create a co-existence between the traditional business publishing approach and the institutional, e-print archive repositories, which are more and more diffused in the scholarly international context. Therefore the framework proposed by the Open Archives Initiative, based on Data and Services Providers and on the Protocol for Metadata Harvesting, has revealed itself to be very suitable to design the system architecture. An OpenURL based resolution system is the most suitable component to support resources cross-referencing in order to assure the integration with the digital libraries context.

Concerning metadata, DAFNE has made evident how much metadata permeates the entire publishing pipeline. Descriptive, relation and rights management are the functional groups of primarily required elements to implement the prototype. Technical, long term preservation and secure (i.e. digital digests and signatures) metadata will be added in the advanced version. The project has been investigating relevant existing metadata standards to define DAFNE metadata set by aggregation of and reference to these vocabularies. Dublin Core and PRISM are the main reference standards for descriptive and relational metadata. PRISM rights and permission metadata have been extended with some elements derived from ODRL model. They make up a minimal set for rights management in order to experiment on-line payment and rights computation functions. By the end of the year a full specification of DAFNE metadata set will be issued. Next year the DAFNE prototype exploitation will allow the consolidation of the final metadata specification.

## References

ArXiv. <http://arXiv.org/>

Berkeley Electronic Press. <http://www.bpress.com>

CogPrints. Cognitive Sciences Eprint Archive.
< http://cogprints.soton.ac.uk/>

DOI. Digital Object Identifier. <http://www.doi.org/>

Eprints.org. University of Southampton.
<http://www.eprints.org>

Ginsparg, P., 2001. Creating a Global Knowledge Network. Conference on Electronic Publishing in Science. Paris, 20 February 2001.
<http://arXiv.org/blurb7pg01unesco.html>

Harnad, S., 1999. Free at Last: The Future of Peer-Reviewed Journals. D-Lib Magazine, 5(12). <http://www.dlib.org/dlib/december99/12harnad.htm>

Iannella, R., 2001. Digital rights management (DRM) architectures.D-Lib Magazine, 7 (6). <http://www.dlib.org/dlib/june01/iannella/06iannella.html>

Iannella, R., 2001a. Open Digital Rights Language (ODRL) Version 1.0. <http://odrl.net/1.0/ODRL-10.pdf>

IETF RFC 2141. URN Syntax. <http://www.ietf.org/rfc/rfc2141.txt>

IETF RFC 2396. Uniform Resource Identifiers (URI): Generic Syntax. <http://www.ietf.org/rfc/rfc2396.txt>

<indecs>, Interoperability of data in e-commerce systems. <http://www.indecs.org>

NCSTRL. Networked Computer Science Technical Reference Library.<http://www.ncstrl.org>

OAI. Open Archives Initiative. <http://www.openarchives.org>

OAIS, 1999. Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-R-1. Red Book. Issue 1. <http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html>

OCLC/RLG, 2001. Preservation Metadata for Digital Objects: a Review of the State of the ART. A White paper by the OCLC/RLG Working Group on Preservation Metadata. < http://www.oclc.org/digital preservation/presmeta_wp.pdf >

Pasqui, V., 2001. DAFNE Deliverable D12: Definizione dell'architettura logica a macroblocchi del sistema complessivo sulla base del flusso dei servizi e delle transazioni da sviluppare nella architettura tecnologica. Draft vers. 1. November 2001. Restricted report.

Pasquinelli, A., 1997. Information technology directions in libraries: a Sun Microsystems white paper: August 1997. <http://www.sun.com/products-n-solutions/edu/libraries/libtechdirection.html>

PRISM 2002, PRISM:Publishing Requirements for Industry Standard Metadata, Feb. 2002. <http://www.prismstandard.org/techdev/primspec11.asp>

Rust, G. and Bide, M., 2000. The <indecs> metadata framework : Principles, models and data dictionary <http://www.indecs.org>

SPARC, 2002. The Case for Institutional repositories: A SPARC Position paper. Prepared by Raym Crow. SPARC. Release 1.0. <http://www.arl.org/sparc/home>

ScholarOne. http://www.ScholarOne.com

Van de Sompel, H. and Beit-Arie, O. Open Linking in the Scholarly Information Environment Using the OpenURL Framework. D-Lib Magazine. 7(3). <http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>

Van de Sompel, H. and Lagoze, C., 2001. The Open Archives Initiative: Building a low-barrier interoperability framework. JCDL 2001. <http://www.openarchives.org/documents/oai.pdf>

XpressTrack. http://xpresstrack.com

XrML. Extensible Rights Markup Language. ContentGuard, Inc. < http://www.xrml.org/>

## Appendix - DAFNE Metadata Element Set (provisional)

The first six functional groups are derived from PRISM specification and the same purpose and meaning is reported for each element.

*General Purpose Descriptive Elements*

***dc:identifier*** Identifier(s) for the resource.
***dc:title*** The name by which the resource is known.
***dc:creator*** The primary creator(s) of the intellectual content of the resource.
***dc:contributor*** Additional contributors to the creation or publication of the resource.
***dc:language*** The principal language of the resource.
***dc:description*** A description of the resource.
***dc:format*** The file format of the resource.
***dc:type*** The style of presentation of the resource's content, such as image vs. sidebar.
***prism:category*** The genre of the resource, such as election results vs. biographies.

*Elements for Provenance Information*

***dc:publisher*** An identifier for the supplier of the resource.
***prism:distributor*** An identifier for the distributor of the resource.
***dc:source*** An identifier for source material for the resource.

*Elements for Time and Date Information*

***prism:creationTime*** Date and time the identified resource was first created.

***prism:modificationTime*** Date and time the resource was last modified.
***prism:publicationTime*** Date and time when the resource is released to the public.
***prism:releaseTime*** Earliest date and time when the resource may be distributed.
***prism:receptionTime*** Date and time when the resource was received on current system.

*Subject Descriptions*

***dc:coverage*** Indicates geographic locations or periods of time that are subjects of theresource.
***dc:subject*** The subject of the resource.
***dc:description*** Prose description of the content of the resource.
***prism:event*** An event referred to in or described by the resource.
***prism:location*** A location referred to in or described by the resource.
***prism:person*** A person referred to in or described by the resource.
***prism:organization*** An organization referred to in or described by the resource.

*Resource Relationships*

***prism:isPartOf*** The described resource is a physical or logical part of the referenced resource.
***prism:hasPart*** The described resource includes the referenced resource either physically or logically.
***prism:isVersionOf*** The described resource is a version, edition, or adaptation of the referenced resource.
***prism:hasVersion*** The described resource has a version, edition, or adaptation, namely, the referenced resource.
***prism:isFormatOf*** The described resource is the same intellectual content of the referenced resource, but presented in another format.
***prism:isTranslationOf*** The described resource is a human-language translation of the referenced resource.
***prism:hasTranslation*** The described resource has been translated into an alternative human-language. The translated version is the referenced resource.

*Rights and Permissions*

***dc:rights*** Container element for specific rights data
***prism:copyright*** A copyright statement for this resource.
***prism:expirationTime*** Time at which the right to reuse expires.
***prism:releaseTime*** Time as which the right to reuse a resource begins, and the resource may be published.
***prism:rightsAgent*** Name, and possibly contact information, for the agency in order to establish contacts and to determine reuse conditions if none specified in the description are applicable.
***prl:geography*** Specifies geographic restrictions.
***prl:industry*** Specifies restrictions on the industry in which the resource may be reused.
***prl:usage*** Specifies ways that the resource may be reused.

*DAFNE elements*

*dafne:usagePermission* Specifies the available access modalities (e.g. display, print).
*dafne:payment C*ontains the following four sub-elements to express payment information.
*dafne:amount* Specifies the cost.
*dafne:currency* Specifies the currency (e.g. €).
*dafne:taxpercent* Specifies the tax percentage (between 0 and 100).
*dafne:code* the tax code (e.g. IVA or VAT).
*dafne:paymentRequirements* Specifies the payment modalities requested (e.g. prepay, post pay, peruse).
*dafne:rightsHolders* Contains the three following sub-elements to specify royalties accounting data.
*dafne:holderId* Specifies the unique identifier of a subject entitled to rights fees.
*dafne:percentage* Specifies the percentage of dafne:amount to be used to compute the fee due.
*dafne:fixedamount* Specifies a fixed amount due for rights.

If different modalities of access exist for the same publication *dafne:paymentRequirements*, *dafne:rightsHolders* and *dafne:payment* are related to different instances of *usagePermission*.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 19-26*


# Metadata in the Context of The European Library Project

Theo van Veen
Koninklijke Bibliotheek, The Netherlands
Theo.vanVeen@kb.nl

Robina Clayphan
British Library, United Kingdom
Robina.Clayphan@bl.uk

## Abstract

*The European Library Project (TEL), sponsored by the European Commission, brings together 10 major European national libraries and library organisations to investigate the technical and policy issues involved in sharing digital resources. The objective of TEL is to set up a co-operative framework which will lead to a system for access to the major national and deposit collections in European national libraries. The scope of the project encompasses publisher relations and business models but this paper focuses on aspects of the more technical work in metadata development and the interoperability testbeds. The use of distributed Z39.50 searching in conjunction with HTTP/XML search functionality based on OAI protocol harvesting is outlined. The metadata development activity, which will result in a TEL application profile based on the Dublin Core Library Application Profile together with collection level description, is discussed. The concept of a metadata registry to allow the controlled evolution of the application profile to be inclusive of other cultural heritage institutions is also introduced.*
**Keywords:** *European Digital Library, Interoperability, Dublin Core Metadata, Application Profiles, Collection Level Description, Search and Retrieve via URLs, SRU.*

## 1. Introduction

The European Library Project (TEL) [7] is partly funded by the European Commission as an accompanying measure under the cultural heritage applications area of *Key Action 3* of the *Information Society Technologies* (IST) research programme.

Co-ordinated by the British Library the project partners are:
Biblioteca Nacional, Portugal (BN)
Biblioteca Nazionale Centrale Firenze, Italy (BNCF)
Conference of European National Librarians (CENL)
Die Deutsche Bibliothek, Germany (DDB)
Helsingin Yliopiston Kirjasto, Finland (HUL)
Istituto Centrale per il Catalogo Unico, Italy (ICCU)
Koninklijke Bibliotheek, The Netherlands (KB)
Narodna in Universzitetna Knjiznica v Ljubljani, Slovenia (NUK)
Swiss National Library, Switzerland (SNL)

The objective of The European Library project is to set up a co-operative framework which will lead to a system for access to major European national and deposit collections. TEL will lay down the policy and develop the technical groundwork for the development of a pan-European digital library that is sustainable over time. The operational system will be implemented once the results of the project are known. Although the focus of the project will be on digital material as provided by the TEL-partners and publishers of digital material, traditional materials are not excluded.

This paper will discuss the development of a metadata model and the development of an interoperability testbed. This testbed will offer distributed searching in the national collections via Z39.50 alongside searching a central index of metadata harvested from other collections via the Open Archives Initiative protocol (OAI) [8]. This central index will be accessible directly via http. The design of the metadata model must enable current functionality and be open to future requirements with regard to the access of collections, digital objects and services.

The combination of distributed searching and central indexing and the use of two major search and retrieve protocols, Z39.50 and http/XML(SRU) - explained later in this paper, make the TEL project unique as similar projects usually use only one or the other access method.

## 2. The workpackages

The TEL-project consists of six workpackages:
1) Relation with publishers
2) Business plans and models
3) Metadata development
4) Interoperability testbeds
5) Dissemination and use
6) Management

This paper will focus on the more technical workpackages: workpackage 3, concerning the metadata development and workpackage 4, concerning the development of the interoperability testbeds.

These workpackages are interdependent: testbeds cannot work without the appropriate metadata and the metadata development needs an operational system for testing and developing the metadata models. It was therefore decided to work on both workpackages in parallel and for each to make use of the other's results in an iterative and incremental way. This meant that at the start of the project, for the http/XML testbed, any metadata format available in XML record syntax could be chosen. The results of the metadata development will be directed towards the operational TEL service and therefore do not have to be available until a later stage in the project. During the course of the project the work on metadata can use the http/XML testbed for the development of ideas and the data model can be brought in line with these ideas.

## 3. Metadata development

The various national libraries and publishers have different descriptive metadata formats. To access these different distributed sets of metadata a common datamodel will be developed. The data model will also support the functionality required within TEL thereby enabling data sharing amongst the TEL partners.

The TEL project aims at consensus building rather than delivering an operational service. Metadata arises from a functional analysis and an operational TEL service will probably reveal more functional requirements than we are currently aware of. The approach being followed is therefore directed towards identifying the functionality we can foresee and defining the metadata needed to support it. The metadata world is becoming more and more complex with an increasing number of standards (such as EAD, MARC, METS, MODS, RDF, DC, ONIX, CIMI, XML), so it will be a big challenge to develop a common datamodel that enables us to find, identify, select and access services from the individual members.

There appear to be two options. One is to convert all the partner's metadata into a single format. An alternative is to develop a metadata model that is generic and that can incorporate multiple metadata standards - the solution to this may be to introduce a TEL metadata-registry system.

At the outset it was agreed to use XML as the record syntax and unqualified Dublin Core as the temporary record schema to enable the test-bed development to proceed. The TEL metadata working group has since concluded that the DC-Library Application Profile (DC-Lib) [3], which is a combination of qualified Dublin Core and other namespaces, would be the best choice as a starting point for the datamodel for the operational TEL service.

## 4. The interoperability testbeds

The work on the interoperability testbeds will initially be focussed on the development of separate testbeds for http/XML and Z39.50, later in the project both will be brought together into one interoperability testbed. For Z39.50 it was agreed to conform to the Bath profile and this conformance will be the subject of testing. For http/XML there is not such a profile. A mapping is needed from user queries to Bath-conformant queries on one hand and the same user queries to http/XML queries on the other hand. A big challenge will eventually be the semantic interoperability between both testbeds.

There are two aspects to the http/XML testbed. First is the development of a mechanism to harvest records from contributing partners and, secondly, the specification and implementation of a protocol to make the data accessible by an external portal. For harvesting it was decided to use the OAI protocol.

At the same time as the specification of a protocol for search and retrieve was underway in TEL, the Z39.50 implementers group were working on the Z39.50 International Next Generation (ZiNG) [10] initiative. Under this umbrella two protocols for Search and Retrieve were initiated: Search and Retrieve via URLs (SRU) and Search and Retrieve via the Web (SRW). SRU uses the URL-GET method for its input; SRW makes use of the Simple Object Access Protocol (SOAP). Both protocols return XML as output and are similar with respect to request parameters, query language and the XML-output. As the original TEL specifications for the http/XML testbed were very close to the SRU specifications, it was decided to follow the SRU-standard for this testbed. It is likely that this will also be used in the final operational TEL-service.

Being one the earliest implementations of the SRU while it is still under development is quite an exciting aspect of the TEL-project.

## 5. Overview of infrastructure

An overview of the infrastructure is shown below. The TEL operational service will be a central portal and/or local portals. Separate portals will be used for

the two testbeds during development. Later in the project these testbeds will be combined to a central portal for the interoperability testing. In the overview this is illustrated by the ellipse around both portals. For the operational TEL service, when the integration of national services is sufficiently stable, the TEL-portal may be mirrored to local portals.

Five partners in the project will provide metadata in XML via the OAI-protocol. These records will be indexed in a central index. The portal will search and retrieve them via the SRU-protocol. The databases of four other partners will be accessible via Z39.50. The metadata will offer links to the digital objects and services. These links will be either direct or indirect via link services using OpenURL's or URN-resolvers. Additional services might be offered, for example multi-linguality (translation of specific subject headings) or thesaurus services allowing the use of search results from a thesaurus database as input for subsequent searches in TEL.

## 6. The approach to metadata development

The first stage in this workpackage consisted of a review of the partner's use of metadata. This "state of the art" report was based on a survey of current practice and desk research. Following that a metadata working group was installed comprising members from each participating library. Analysis of the state of art review resulted in the decision to define the metadata requirements by analysing the functionality required for TEL, and then determining what metadata elements were needed to fulfil those requirements.

### 6.1. State of the art review

Analysis of the responses to the metadata questionnaire produced five main conclusions:

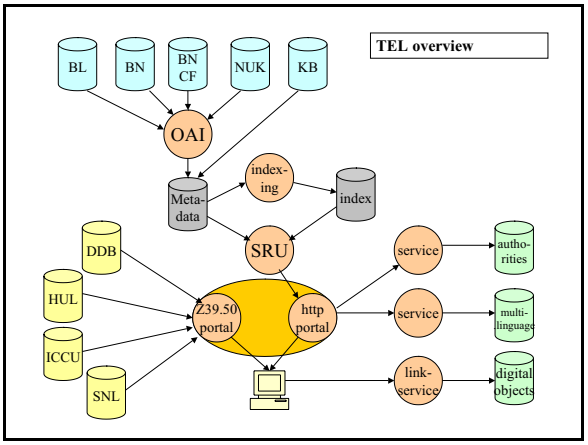



**Figure 1. Overview of TEL infrastructure**

1. There is no consensus between the partners about categories of metadata. Partners have interpreted the categories differently according to the scope and purpose of their implementations. This is illustrated in the following diagram, which shows how the partners defined different categories of metadata.

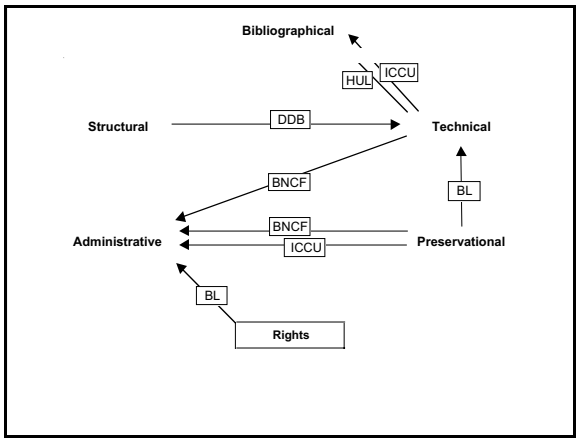



**Figure 2. Overview of the differences in terminology of metadata**

2. Libraries need to share knowledge on the creation of metadata, especially for collections that will be described in the future.
3. The absence of a common bibliographic format makes simultaneous access to metadata from different partners difficult. Formats in use are:
   - MARC21
   - Finmarc [1]
   - Dublin Core
   - UNIMARC
   - Pica3
   - PicaPlus
   - COMARC
   - Custom built datamodels

   The custom built datamodels in particular cause a problem: there are many of them and there are no generally available mappings for them. It is expected that the use of Dublin Core (or DC-Lib) will make it easier to develop consistent resource discovery services.
4. There is not yet one linking mechanism or resolution service used by all partners. Research is needed on the use of metadata for linking to external resources by means of URNs, PURLs or OpenURL.
5. There is uncertainty about the eventual contents of TEL and a need to be aware of the danger of an unbalanced service that may render small collec-

[1] Finmarc has in the meantime been replaced by MARC21.

tions invisible amongst complete union catalogues.

### 6.2. Developing the TEL datamodel

It was considered very desirable to adopt existing metadata standards where possible. The DC-Library Application Profile was identified as the most obvious candidate to form the basis of the TEL datamodel. The general approach has been to define the functionality needed to underpin the services and publication types as currently envisaged on the one hand and identify the metadata required to enable that functionality on the other. Special attention will be paid to digital collections and collection level descriptions (CLD).

The functional requirements could then be analysed against DC-Lib to see what gaps existed in the element set. From these results we can determine whether DC-Lib is sufficient for TEL or whether it will be necessary to define a TEL Application Profile which will incorporate DC-Lib with additional elements. Another possibility is to request the DC-Lib drafting committee to incorporate the additional metadata elements into DC-Lib, but the timescales of the project may not allow this.

Finally we need to determine the best mapping between the TEL Application Profile and the partners various metadata schemas and bibliographic formats. Desk research and experimentation with actual data from the different partners will determine how to implement the application profile in XML. The results will become part of the TEL metadata handbook, which will be made available on the web to facilitate the introduction of new collections to TEL.

It is not envisaged that TEL will stop at the end of the project but will continue to evolve afterwards. The impact of this is that the development effort will not be solely focussed on the TEL test-bed, but a more generic approach will be followed. This will allow future TEL functionality to be taken into account. It also raises the possibility of another approach to metadata specification. This option is to create a TEL registry of metadata that would allow the addition of new metadata elements that are in use by the partners. This possibility is discussed later in this paper but at the time of writing this approach has not been fully discussed within the project.

## 7. Functionality and services

The analysis of functionality and services was the result of desk research and included mapping functions to metadata elements. Functions considered relevant for TEL were put on the horizontal axis of a matrix and the metadata elements were put on the vertical axis. Functions refer to TEL as a whole and not solely to the TEL-portal. The elements were those from draft Library Application Profile of 2001-10-12. The complete overview is contained in a project report which is not publicly available at the time of writing. The mapping was intended to highlight any gaps where a function could not be supported by the available metadata.

The main functions are:

- *Search and resource discovery*
  This is fundamental functionality. Most metadata elements contribute to this function.
- *Record retrieval*
  Record retrieval follows from a search and also plays a role in harvesting and indexing metadata. Metadata elements for the identification of the original record is required as is specification of the record format.
- *Identification of resources*
  Needed to find and access resources. All elements used for referencing may play a role.
- *Description*
  Many metadata elements help the user in a decision to obtain the object.
- *Linking services*
  Linking services help locating objects or services. In many case these are resolution services. All metadata elements that play a role in dynamic linking are relevant.
- *Multilinguality service*
  Envisaged as a service translating user input or returned metadata into different languages to create new queries or search terms. Most textual metadata may play a role in this.
- *Thesaurus service*
  Envisaged as a service to find main entries for subjects and classification from user input or returned metadata to create new queries. Textual metadata or classification codes may play a role.
- *Collection level services*
  The functionality that helps to find and identify collections, link to those collections or broadcast queries to distributed collections.
- *Authorisation*
  Access may depend on the service, type of publication, year of publication, publisher, the user etc. Restrictions are indicated by terms and conditions or access rights.
- *Administration*
  Functions that keep track of usage, based on, for example, subject or publisher.
- *Hard and software requirements*
  Specific metadata to inform users of the requirements on their workstations or detect whether the users workstation is capable of accessing a publication type. Especially when preservation activities play a role.
- *Navigation*
  This functionality concerns linking related metadata records by dynamic linking, for example tracking hierarchical relationships like journal-issue-article or expression-work-manifestation etc.

**Tabel 1. Mapping of metadata elements to functions by their usage**

| Element | Qualifier/ Scheme/ Role | Search/resource discovery | Retrieval of metadata | Identification | Description | Link service | Multilinguality | Thesaurus service | Collection level | Authorisation | Administration | Hard and software. | Navigation | Copy cataloguing | Miscellaneous | Comment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Record Identifie | Any | | X | X | | X | | | | X | | | | | X | To identify the metadata record |
| Identifier, Source, Relation | Base-URL | | X | ? | | X | | | X | | | | X | | X | Encoding scheme for a UR with variable query |
| Identifier, Source Relation | URN | | | X | | X | | | X | | | | X | | X | Encoding scheme foURNs |
| Identifier | PURL | | | X | | X | | | X | | | | X | | X | Encoding schema for persistent URLs |
| Identifier, Source, Relation | OpenURL | | | X | | X | | | X | | | | X | | X | Encoding scheme for queri with a variable baseURL |
| Collection level description Schema | All | X | | | X | | | | X | X | X | | X | ? | ? | Possibly RSLP or the DC schema. |

- *Copy cataloguing*
  Metadata may be re-used by other libraries for cataloguing.
- *Miscellaneous*
  Off-line ordering, ILL and other services. Mostly accessed directly via URLs. Link services are anticipated for TEL. Metadata regarding holding or item information and identification of the original metadata record are most important for TEL.

Mapping the functions and services listed above to the draft DC-Lib of 2001-10-12 some functions can be seen to need additional metadata elements or encoding schemes. This is shown in the table below.

The above table shows the metadata elements or qualifiers that are not present in DC-Lib, but will be needed for some of the required functionality. This will be discussed below. It should be noted that there will be many more metadata elements that will be useful or even necessary to search and access the digital objects from specific collections. To identify these metadata elements the specific collections will have to be examined. How these can be handled is discussed in the Registry section of this paper.

One aspect that still needs special attention, but which is not yet covered in this paper, is the sophistication of search functionality based on the semantic relations between metadata – as described in "The ABC Ontology and Model" [5] for example. The complexity and the human effort needed to create records that support queries based on these more complex semantics are expected to be rather high. This aspect of the functionality will therefore be addressed separately from the basic questions regarding which metadata are needed.

A further aspect of the relationship between functionality and metadata concerns which fields (access points or indexes) that can be searched. All metadata elements are – as long is it is reasonable – implicitly considered to be searchable and a one to one relation between search field and metadata element is assumed.

### 7.1. Metadata for linking

Most of the metadata elements, that need special attention, have to do with linking. Identification of the metadata record is relevant for TEL in order to maintain the reference to the original records for harvesting purposes and when record identifications are used in dynamic URLs (linking by reference).

TEL has to deal with metadata that should, as far as possible, be independent of the publication or service that is described in the metadata record. In other words, the portal should not have interpret the content of elements but simply act on the rules governing the type of metadata. For the identifier element these rules will generally be different for OpenURL, URLs, URNs and PURLs. The dynamic and context sensitive creation of links in which special link services or resolution services will be involved, will be different for these types of identifier. Using only URI as identifier encoding scheme for linking in DC-Lib will therefore not be sufficient. This also concerns the source and relation elements.

A special type of encoding scheme is the base-URL. This base-URL identifies a collection and will be used in generating URLs representing queries into such a collection (deep linking). This is different from the conventional URL for accessing the website of a collection.

### 7.2. Collection level descriptions

Collection level descriptions have a place in the TEL metadata set as TEL is essentially a collection of collections. To describe collections, metadata elements are needed that are not used in the description of conventional publication types. Any aggregation of objects could be considered as a collection and a collection can simply be considered as a top level container of records. An important aspect of collections is the way they are accessed: some collections can be searched individually and others are simply a static website.

In TEL there are two different ways to look at collections: 1) they can be considered to be a publication like any other, but being of a specific type, 2) they can be considered as an aggregation of publications. In the latter case a collection may be a target for distributed queries. These two aspects of collections give rise to a potentially very powerful future functionality: they allow the user to find collections as the result of a search and then select these collections as the list of targets for a next – more precise – distributed search.

The importance of collection level descriptions is such that it justifies a complete new set of metadata elements. The resource implications of discussing each individual metadata element for a collection level description within TEL would be onerous. In line with the principle adopting existing (or developing) standards, TEL will utilise an existing CLD

schema such as that developed by RSLP [6]. The DC Collections working group [2] is also considering the RSLP schema. After further work it is anticipated that TEL will include the complete schema for collection level descriptions in its own metadata set, therefore in the functionality matrix no individual metadata elements for collection level elements are shown.

## 8. TEL metadata registry and metadata formats

Although DC-Lib was identified as a valuable starting point for the TEL application profile it does not contain all the elements that TEL will need. Even if it would suffice for now it will not be sufficient in the future when new functionality is introduced. We therefore need to create a TEL application profile which may contain elements that are not part of a DCMI namespace at the moment.

As seen now, The European Library project is a system for access to all types of collections and materials owned by the European Libraries. In the future it may be opened to other types of memory institutions and, if so, the issue of semantic interoperability will become an important aspect of the development. The flexible structure of Dublin Core and the different sectoral interpretations of how to describe a digital object could be an obstacle to interoperability. In Dublin Core there are no rules of the kind we employ in libraries for how the values in metadata are constructed and a unitary search does therefore not guarantee the localisation of all types of digital objects.

In the creation of a TEL profile based on DC-Lib it is also important to define a model and an ontology as a starting point for the development of vocabularies relating to different applications in the Cultural Heritage sector. Libraries own and catalogue materials that are also owned and catalogued in different types of institution (Archives and Museums for example). It is important to define the correspondence between terms, functions and concepts in the systems describing that material.

There are may be several ways of dealing with this:

1. Promote a common standard schema, independent of what different data providers are using internally. This would entail the definition of one comprehensive mapping table using which all information providers could convert their metadata to a single TEL-schema.
2. Introduce a TEL metadata registry that contains metadata from existing metadata standards, but which can be extended with local metadata. The Library application profile will be the "main" entry, but as soon as new metadata are introduced for which there is no existing element in DC-Lib application profile, or other profiles accepted by TEL, then the introduction of new elements would be allowed. In this context, the TEL Registry is seen as a system that facilitates the procedures involved in allowing the TEL application profile to evolve in line with increased functionality and extension to different types of institution. In this it is slightly different from the concept of a registry in the sense it is currently used in DCMI [9].
3. Use the TEL indexes mainly for the first FRBR [4] objective i. e. to find all resources sharing the same index entry and – for the other objectives – the user will be redirected towards the real catalogs. In this context, Dublin Core presents itself as a metadata pidgin for digital tourists who must find their way in this linguistically diverse landscape [1].

The first option is preferred but we may need the second option to realise the first one: the registry will define a common standard schema, but building the schema in a decentralised and incremental way is enabled. Data providers would be allowed to add new metadata elements but at the same time all providers can monitor the developing schema and raise objections to inappropriate metadata elements. The third option is a last resort option for very specific cases to provide metadata elements from the original record for which there are no corresponding metadata elements in the registry.

The TEL portal would use the metadata elements from this schema/registry for such actions as display, translate, generate a link or generate a new search. This TEL registry would therefore contain information additional to the basic ontology on how TEL will handle these metadata.

The registry will contain at least:

- element name
- name space
- originating metadata standard
- labels for presentation in different languages
- flag to indicate that it should be presented in the full presentation
- flag to indicate whether the element should be used as clickable link for searching
- element name that it maps to (in case it comes unconverted from another metadata standard)

This list may grow in the future as the usage of different metadata elements in practical situations is extended. When new collections from national libraries enter the TEL-system the flexibility of such a registry will facilitate compliance to the TEL system.

## 9. Implementation

In the figure below an overview is shown of the steps involved in exchanging metadata.

There are several formats involved here, that can also differ per project partner. That is:

1) XML, this can conform to any element set but the

Library application profile is preferred
2) TEL-XML, which is defined by the TEL-project (DC-Lib)
3) MARC
4) HTML meant for the user

For Z39.50 the local data will be provided via the local Z39.50 servers in TEL-XML but in case that is not supported, MARC is also accepted. For the http/XML sources the SRU server will perform some conversions if the original data is not conform DC-Lib.

In the TEL portal the final conversions will take place to provide the user presentation. This conversion makes use of the metadata registry in which it is defined how the metadata should be processed.

In the above overview there is an additional Z39.50 gateway, which could transform the Z39.50 output records to TEL-XML. This is not implemented or even planned but is just an extra possibility that would allow the portal to deal with one single XML format instead of XML and MARC.

As there were no portals available that supported distributed search conforming to the SRU protocol at the start of the project, a test-portal was developed for this purpose. This test-portal is based on XSL and javascript and runs locally in the browser. It was found to be quite convenient to have everything at one place during the test phase and therefore the metadata registry is initially implemented in tables in the same javascript that makes up the local portal.

## 10. Handbook

A TEL metadata handbook will be made available to help the partners in submitting collections and new metadata elements to TEL. This handbook will contain information on metadata mappings, conversion schemes, standards, relevant links and TEL requirements with respect to metadata. It will help the TEL partners to define an ontology in developing an integrated glossary (or specific vocabularies) related to the different systems and services.

## 11. Conclusion

Projects working to limited timescales in the rapidly changing world of digital resource sharing have to adopt a pragmatic approach to development activity. The parallel development of the testbeds and the metadata schema has allowed work to proceed on both these aspects of TEL without delay to either activity and will allow the development of each to incorporate the findings of the other. Adoption of emerging standards minimises the need to spend time developing customised solutions and should assist wider interoperability as well as adding an element of future-proofing to the resulting system. In

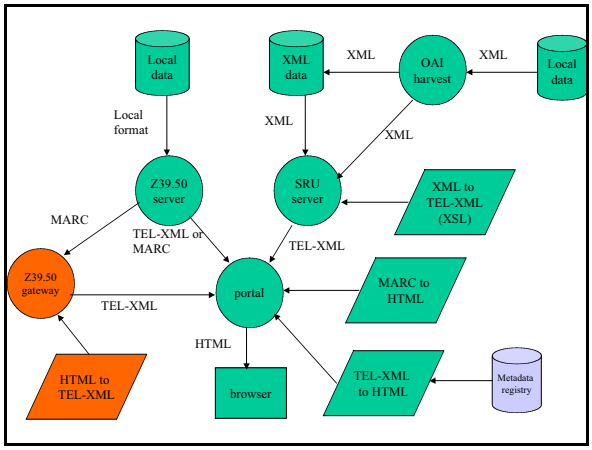



**Figure 3. Metadata formats in different processing stages**

using current developments in Z39.50 and Dublin Core TEL is confident that it has chosen well-founded namespaces with which to work in this respect. Resource sharing across nations necessitates working with heterogeneous collections – the use of OAI harvesting and central indexing of XML records facilitates integrated access to these. Looking to the future means developing systems that are open to evolution – the proposed metadata registry is an attempt to build more openness into the system.

Although the current TEL project will not result in a fully operational European Library system, the results of the project will constitute the groundwork on which that ambitious vision can be realized.

## References

[1] Baker, Thomas, A Grammar of Dublin Core, 2000, http://www.dlib.org/dlib/october00/baker/10baker.hml

[2] DC Collections Working Group http://dublincore.org/groups/collections/

[3] DCMI Libraries Working Group http://www.dublincore.org/groups/libraries/

[4] Functional Requirements for Bibliographic Records, International Federation of Library Associations and Institutions, March 1998, http://www.ifla.org/VII/s13/frbr/frbr.pdf

[5] Lagoze, Carl and Hunter Jane, The ABC Ontology Model, in Journal of Digital Information, 2001 http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/

[6] Powell, A., RSLP Collection Description Schema http://www.ukoln.ac.uk/metadata/rslp/schema/

[7] The European Library
http://www.europeanlibrary.org/

[8] The Open Archive Protocol http://www.openarchives.org/OAI/openarchivesprotocol.htm

[9] The Open Metadata Registry http://wip.dublincore.org:8080/registry/Registry

[10] Z39.50 international Next Generation
http://www.loc.gov/z3950/agency/zing/

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 27-34*


# Archon - A Digital Library that Federates Physics Collections

K. Maly, M. Zubair, M. Nelson, X. Liu, H. Anan, J. Gao, J. Tang, Y. Zhao
Computer Science Department
Old Dominion University
Norfolk, Virginia, USA
{maly,zubair,mln,liu_x,anan,gao_j,tang_j,yzhao}@cs.odu.edu

## Abstract

*Archon is a federation of physics collections with varying degrees of metadata richness. Archon uses the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to harvest metadata from distributed archives. The architecture of Archon is largely based on another OAI-PMH digital library: Arc, a cross archive search service. However, Archon provides some new services that are specifically tailored for the physics community. Of these services we will discuss approaches we used to search and browse equations and formulae and a citation linking service for arXiv and American Physical Society (APS) archives.*

## 1. Introduction

Archon is a federation of physics digital libraries. Archon is a direct extension of the Arc digital library [13]. Its architecture provides the following basic services: a storage service for the metadata of collected archives; a harvester service to collect data from other digital libraries using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [10]; a search and discovery service; and a data provider service to expose the collected metadata to other OAI harvesters. However, for Archon we have developed services especially for physics collections based on metadata available from the participating archives that go beyond the required (by the OAI-PMH) unqualified Dublin Core (DC) [22]. For example, we provide a service to allow searching on equations embedded in the metadata. Currently this service is based on LaTeX [11] representation of the equations (due to the nature of archives used), but we plan to include MathML [8] representations in the near future. We also use context-based data to search for equations related to specific keywords or subjects. By intelligent template matching, a cross-archive citation service has been developed to integrate heterogeneous collections into one unified linking environment.

## 2. Overview of Archon Services

The Archon architecture is based on the Java Servlets-based search service that was developed for Arc and earlier for the Joint Training, Analysis and Simulation Center (JTASC) [16]. This architecture is platform independent and can work with any web server (Figure 1). Moreover, the changes required to work with different databases are minimal.

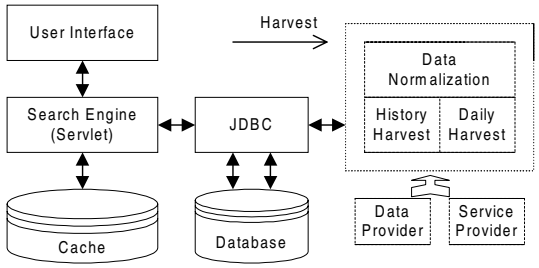



**Figure 1. Overall Architecture**

### 2.1 Search Service

The search server is implemented using Java Servlets (Figure 2). The session manager maintains one session per user per query. It is responsible for creating new sessions for new queries (or for queries for which a session has expired). Sessions are used because queries can return more results (hits) than can be displayed on one page, Caching results makes browsing through the hits faster. The session manager receives two types of requests from the client: either a request to process a new query (search); or a request to retrieve another page of results for a previously submitted query (browsing). For a search request, the session manager

calls the index searcher that formulates a query (based on the search parameters) and submits it to the database server (using JDBC) then retrieves the search results. The session manager then calls the result displayer to display the first page. For a browsing request, the session manager checks the existence of a previous session (sessions expire after a specific time of inactivity). If an expired session is referenced, a new session is created, the search re-executed, and the required page displayed. In the case where the previous session still exists, the required page is displayed based on the cached data (which may require additional access to the database).

### 2.2 Storage Service

The OAI-PMH uses unqualified Dublin Core as the default metadata set. Currently, Archon services are implemented based on the data provided in the DC fields, but in the prototype implementation we are already using richer metadata sets. All DC attributes are saved in the database as separate fields. The archive name and set information are also treated as separate fields in the database for supporting search and browse functionality. In order to improve system efficiency, most fields are indexed using full-text properties of the database, such as the Oracle InterMedia Server [18] and MySQL full-text search [9]. The search engine communicates with the database using JDBC [20] and Connection Pool [17].

### 2.3 Harvester

Similar to a web crawler, the Archon harvester (same as the Arc harvester) traverses the list of data providers and harvests metadata from them. Unlike a web crawler, the Archon harvester performs metadata normalization, and exploits the incremental, selective harvesting defined by the OAI-PMH. Data providers are different in data volume, partition definition, service implementation quality, and network connection quality: all these factors influence the harvesting procedure. Historical and newly published data

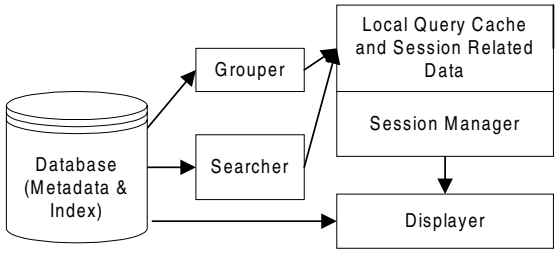



**Figure 2. Search Server Implementation**

harvesting have different requirements. When a service provider harvests a data provider for the first time, all past data (historical data) needs to be harvested, followed by periodic harvesting to keep the data current. Historical data harvests are high-volume and more stable. The harvesting process can run once, or, as is usually preferred by large archives, as a sequence of chunk-based harvests to reduce data provider overhead. To harvest newly published data, data size is not the major problem but the scheduler must be able to harvest new data as soon as possible and guarantee completeness – even if data providers provide incomplete data for the current date. The OAI-PMH provides flexibility in choosing the harvesting strategy; theoretically, one data provider can be harvested in one simple transaction, or one is harvested as many times as the number of records in its collection. But in reality only a subset of this range is possible; choosing an appropriate harvesting method has not yet been made into a formal process. We define four harvesting types:

1. bulk-harvest of historical data
2. bulk-harvest of new data
3. one-by-one-harvest of historical data
4. one-by-one-harvest of new data

Bulk harvesting is ideal because of its simplicity for both the service provider and data provider. It collects the entire data set through a single http connection, thus avoiding the overhead of multiple network connections. However, bulk harvesting has two problems. First, the data provider may not implement the optional resumptionToken flow control mechanism of the OAI-PMH, and thus may not be able to correctly process large (but partial) data requests. Secondly, XML syntax errors and character-encoding problems are surprisingly common and can invalidate entire large data sets. A discussion of general issues regarding metadata variability in OAI-PMH harvesting can be found in Liu, et al. [14].

One-by-one harvesting is used when bulk harvesting is infeasible. However, this approach imposes significant network traffic overhead for both service and data providers since every document requires a separate http connection. The default harvesting method for every data provider begins as bulk harvest. We keep track of all harvesting transactions and if errors are reported, we determine the cause and manually tune the best harvesting approach for that data provider.

The Arc harvester is implemented as a Java application. At the initialization stage, it reads the system configuration file, which includes properties such as user-agent name, interval between harvests, data provider URL, and harvesting method. The harvester then starts a scheduler, which

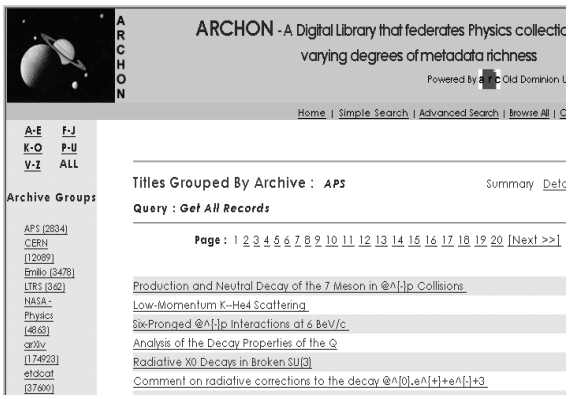



**Figure 3. Archon Interface for Searching**

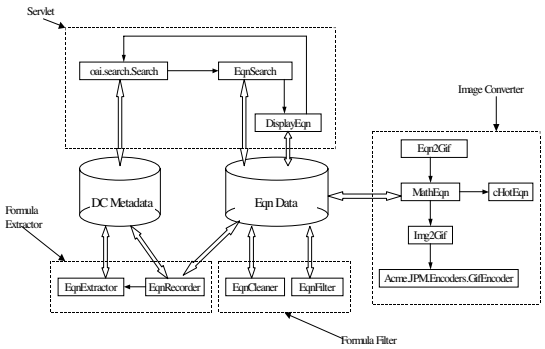



**Figure 4. Equation Search and Display Service Architecture**

periodically checks and starts the appropriate task. Some archives such as Emilio [5] were not OAI-PMH compliant. To overcome this problem, we created a gateway that crawls the Emilio web site and acts as a data provider to provide metadata that is harvested into Archon (Figure 3).

### 2.4 Data Provider Service

The data provider service manages OAI-PMH requests to Archon and allows Archon to act as an aggregator for the metadata contents it harvested from other digital libraries.

## 3. Equations-Based Search

In Archon, many metadata records contain equations in LaTeX and other formats. These equations are harvested as text format and not easy for users to browse and view. It is a value-added service to search equations by traditional text query but present it in a user-friendly way (e.g GIF file). By this method we build virtual metadata (images) over the original flat text metadata. Issues that where addressed to enable Archon to search and browse equations include:

1. Rendering of equations and embedding them into the HTML display.
2. Identifying equations inside the metadata.
3. Filtering common meaningless equations (such as a single $n$) and incomplete equations.
4. Equation storage.

### 3.1 Rendering of Equations

Most of the equations available on Archon are written in LaTeX. However, viewing enocoded LaTeX equation is not as intuitive as viewing the equations themselves, so it is useful to provide a visual tool to view the equations. There are several alternatives to display equations in a HTML page. One alternative is to represent equations using HTML tags. This is an appropriate choice only for simple expressions; using this method severely limits what can be displayed with the usual notation. A browser may not be able to properly display some special symbols, such as integral or summation symbols or Greek characters. The alternative we chose is to write a program to convert the LaTeX equations into an image and embed it inside the HTML page. We implemented this tool as a Java applet.

### 3.2 Identifying Equations

LaTeX equations have special characters (such as $) that mark the start and end of LaTeX strings. However, the presence of these symbols does not automatically indicate the presence of equation. Moreover, an equation can be written as a sequence of LaTeX strings instead of as a whole LaTeX string. This is why we implemented a simple state machine based program to identify equations. Some of the rules used in this state machine are:

1. Isolate the unpaired '$' symbol;
2. Glue the small pieces together into the whole formula;
3. Check the close neighbors (both ends) of a LaTeX string to obtain a complete equation.

### 3.3 Filtering Equations

Despite our progress to date, there are many situations which cannot be solved by the methods described above,

because it is impossible to distinguish if a string is a part of formula when it is not quoted with '$' symbols. We have some "broken" formulas due to this reason. We worked around these limitations by filtering those formulae out. We established a "rule book" where every rule is a pattern of the regular expression which describes what kind of LaTeX string is going to be dropped. Every collected LaTeX string is checked against the rules and ay matching LaTeX strings are removed.

Furthermore, there are also some formulae with 'illegal' LaTeX symbols. Some of these 'illegal' symbols are misspellings, such as a missing space or mistaken use of the backslash symbol. Some of these symbols are user defined. A general-purpose LaTeX string parser cannot properly handle them. All of these will cause a blank image or a formula with missing parts, because the image converter cannot pick up the corresponding display element for it. To solve this problem, each extracted LaTeX string is screened and strings having 'illegal' symbols are dropped.

### 3.4 Equation Storage

For fast browsing, we stored the extracted equation in a relational database. Figure 4 shows the schematic class diagram that shows the relationships between the classes and the relationships between the classes and the database.

Overall, we provide a novel search function, search with equation, to our digital library. To realize this function, LaTeX strings that are used to express equations are extracted from the metadata records. The extracted LaTeX strings are filtered and cleaned to eliminate errors and illegal symbols. Then the clean LaTeX strings are converted into GIF images. We have provided three search alternatives for the user in the search interface Figure 5.

1. Search for the LaTeX string directly.
2. Display a list of all equations and the user can select an equation visually.
3. Search for equations by subject or abstract keywords.

For example, when a user types in a word such as 'Newton' into the 'abstract 'field in Figure 5, we will present to the user all images of formulae that occur in the abstract of papers that contain the keyword 'Newton'. Once a user has selected a subject entry in the box shown in Figure 5, we again display all formulae that occur in papers categorized as having that subject. Finally, by clicking on the formula such as shown in Figure 6, users will receive all the records related to this formula.

At this point we have completed this service for arXiv and are in the process to include the other archives shown in Figure 3. Our approach is to convert all local representation to LaTeX and then use the currently implemented scheme.

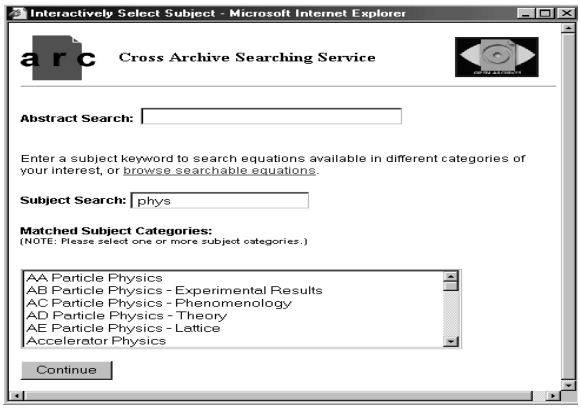



**Figure 5. Formula Search Interface**

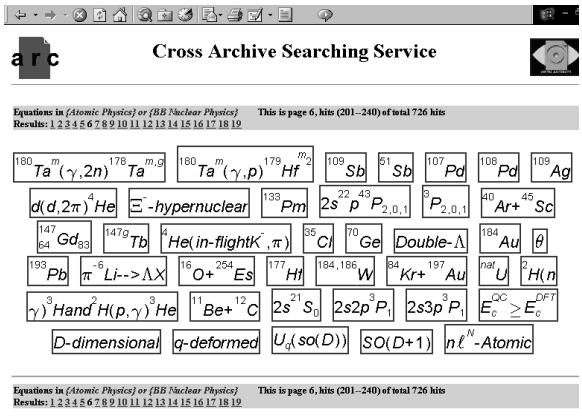



**Figure 6. Formula Search Result Page**

For instance, APS will export its equation in the metadata records in MathML through OAI-PMH's parallel metadata harvesting scheme and we will translate them to LaTeX and store them in our database. CERN already uses LaTeX so it is only a matter of time to use their metadata records.

## 4. Reference Linking Service

The reference linking service provides a convenient method to access references in a document. It provides the references information for a document as well as links from the references to their corresponding documents. There are several kinds of reference linking services. One method is to provide reference links within a controlled collection, such as the Open Journal Project [7]. The feature of our reference linking service is to provide reference-linking service among several collections, the membership of which is subject to change.. The service architecture is shown in Figure 7. In addition to providing reference service for Archon users, we will consider extending our approach for:

1. OAI Citation Provider: Implementing an OAI layer to let other service providers to harvest the citation information from our collections.
2. Public Cross-linking Service: Users can get the reference information by issuing an OpenURL request [21].

The following sub-sections describe our approach in implementing reference linking along with a number of issues that we addressed.

### 4.1 Obtaining Reference information

In order to acquire reference information, we divided the sources into three categories:

**OAI-Compliant Data Provider** Some data providers, such as APS and CiteBase (http://citebase.eprints.org), provide reference information in their specific metadata formats. CiteBase extracts citation information from LaTeX source files in arXiv. In this case, we harvest reference information directly.

**Online Citation Service** Some data providers, such as arXiv, provide online Citation Service. When a user searches an article in arXiv, he/she can press "reference" link to get the reference information about this article. In this case, we have two choices: use a gateway to make it OAI-compliant, or issue HTTP requests to get HTML files and process HTML files directly.

**Articles** It is possible that there is neither harvestable reference information nor an online citation service available. In this case, one may directly extract the reference information from the article's text. CiteSeer [12] and CiteBase have used this approach.

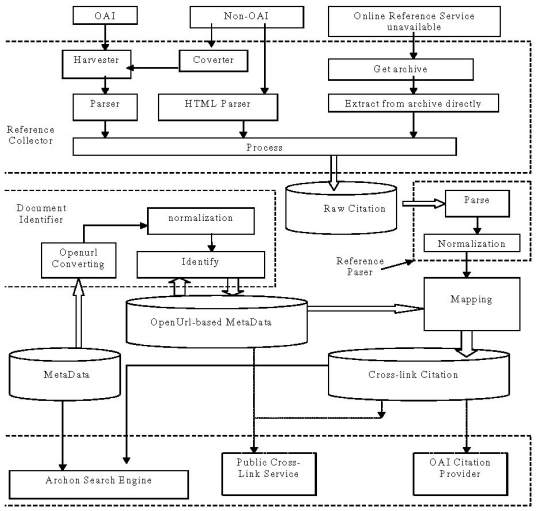



**Figure 7. Service Architecture for Reference Linking in Archon**

### 4.2 Internal Citation Format and Citation Parsing

Since documents and citations are collected from heterogeneous sources and formats, they must be integrated into one unified format for processing. We also consider re-exposing citation information and supporting third-party linking in a standard way during design. Several formats have been studied such as Academic Metadata Format [3] and DC-Citation standard [1]. However, The DC-Citation standard is still under development, and it only addresses journal citation information so far.. We chose the OpenURL [21] metadata sets for four reasons:

1. OpenURL is expressive and accurate in identifying a document.
2. OpenURL has been widely supported and evolves in the context of its NISO standardization.
3. We plan to extend our service to support OpenURL in the future.
4. The possible converge of DC and OpenURL as discussed in [19].

The heterogeneous citation data are converted to the OpenURL metadata set. For example, APS citations are expressed in a XML format that makes it easy for parsing. An example of an APS citation is:
$< citation cid = "C" >$

$< ref >$
$< inbook >$
$< refauth > Langley < /refauth >$
$< booktitle > ReportonMt.WhitneyExpedition$
$< /booktitle >, Profess.Papers,$
$< publisher > U.S.SignalService < /publisher >$
$< volume > XV < /volume >$
$< /inbook >$
$< /ref >$
$< /citation >$

For CiteBase references, the citation information is stored in a semi-structured string that requires heuristics to parse. An example of a CiteBase citation is:
$< relation.References >$
$H.F.Fongetal.Phys.Rev.Lett., 75 : 316, 1995$
$< /relation.References >$

In this case, we implemented a simple state machine based program to parse the citation information. The state machine tries to match the citation against several predefined patterns. Some patterns are:

1. Creators, article title, journal short title, volume, issue, start page, year
2. Creators, article title, journal short title, volume, issue, start page-end page, year
3. Creators, journal short title, volume, issue (year), start page
4. Creators year journal short title, volume, issue, start page

Since no uniform format is defined, normalization and other heuristic processing are necessary. For example, our heuristic algorithm will identify "Phys. Lett.", "Phys. Lett. B 15", "Phys. Lett 15B", "Phys. Letter B 15" as the same journal.

### 4.3 Match between citations and documents

The OpenURL metadata sets almost cover every field that is necessary to identify a document. But document and citation only use a subset of these fields. It is possible that some documents use only the first author and article title while others use journal title, volume number and start page. In our approach, we use multiple rules to match citations and documents based on what kind of the information is present.

Despite our effort, there are many cases that a reference fails to match any document in our collections. There are two possibilities: the referred document does not exist in our collection or the referred document exists in our collection but the matching algorithm failed to find the document

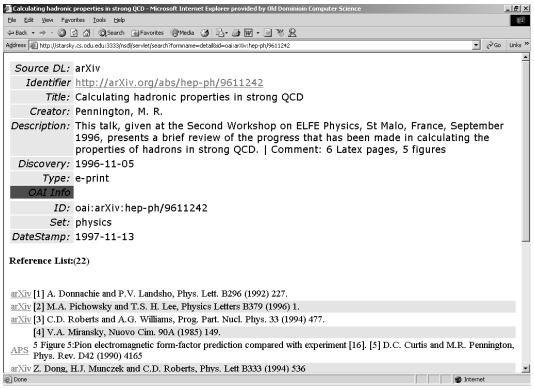



**Figure 8. Reference Display in Archon**

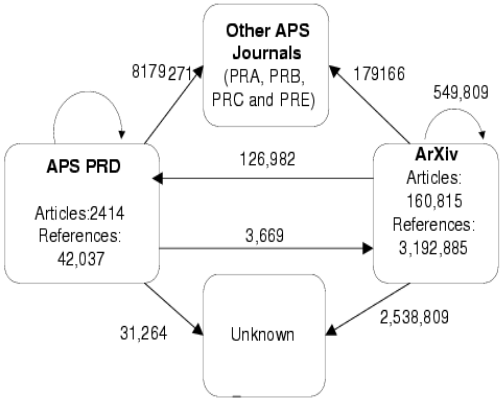



**Figure 9. Initial Result of References Processing in Archon**

Our approach is to compute the similarity between the citation and documents to find possible links. The possible links are presented to the user if the similarity is larger than a pre-defined threshold. The advantage of this approach is that it gives users some possible links in which users may be interested.

Since Archon harvests from various data sources, the same document may exist in more than one source. For example, there is significant overlap between arXiv and APS. To address this problem, we developed algorithms to detect duplicate documents. The duplication is presented the user and we leave the user to select the appropriate copy.

Figure 8 shows the user interface of reference information. Figure 9 shows number of references identified between APS (a subset of the journal Physical Review D) and arXiv. The number of unidentified or unknown references

is currently large, but we are addressing this number by applying more intelligent normalization techniques to identify references.

## 5. Discussion and Future Work

Future work will include updating the Archon harvester and data provider to be compliant with OAI-PMH 2.0, which was released in June 2002 and will run concurrently with OAI-PMH 1.1 until the end of 2002. At that time, OAI-PMH 2.0 will be the only version officially sanctioned by the OAI. Fortunately, OAI-PMH 2.0 represents an incremental evolution of version 1.1, so conversion will not be difficult. Usage of unqualified DC as a common metadata format in OAI-PMH proves to be very helpful for building a quick prototype. However, richer metadata formats are essential for building a richer service. All of the data providers harvested by Archon support metadata formats richer than unqualified DC. Specific parser and citation extraction algorithm have been developed for each of these metadata formats. We consider a standard and rich metadata format for scholarly communication is essential for building richer service over a large number of heterogeneous data providers.

We also plan to continue to refine the equation and citation services. For the equations, we plan to define categories of equations and allow "fielded" searching within those categories of equations. We believe this will increase the precision of equation-based searching. We also created some interfaces for equation search and we are planning to adapt these interfaces to be easier to use from the user point of view. For the citation linking service, we intend to increase the accuracy of our citation parsing and more fully support the OpenURL reference linking framework.

In summary, we created Archon, a digital library for physics. We added services for easier search and browsing of archives as well as their related documents. Our collection includes several OAI-PMH compliant repositories such as arXiv and non OAI-PMH compliant repositories such as Emilio. Other projects like CiteBase, Cyclades [4] and Torii [2] also provide value-added service for physical collections and we plan to compare these services and explore the possibility of cross service linking. At this point it is only our contention that adding equation based search and full cross-linking across all participating archives is a valuable service. In the months to come we will perform user testing to see if these service are welcomed by the physics community. Our prototype implementation has implemented standard ways to ingest metadata of different degree of sophistication and representation and make use of them in a meaningful way.

## References

[1] Apps, A. (2002) A Journal Article Bibliographic Citation Dublin Core Structured Value. Available at http://epub.mimas.ac.uk/DC/citdcsv.html

[2] Bertocco, S. (2001). Torii, an Open Portal over Open Archives, HEP Libraries Webzine, 1(4). Available at http://library.cern.ch/HEPLW/4/papers/4/

[3] Tim D. Brody, Zhuoan Jiao, Thomas Krichel & Simeon M. Warner (2001) Syntax and Vocabulary of the Academic Metadata Format. Available at http://amf.openlib.org/doc/ebisu.html

[4] Cyclades project. http://www.ercim.org/cyclades/

[5] Emilio. AIP Emilio Segr Visual Archives, American Institute of Physics. Available at ¡http://www.aip.org/history/esva/use.htm¿

[6] Harnad, S. & Carr, L. (2000). Integrating, navigating and analyzing open eprint archives through open citation linking (the OpCit project). Current Science Online, 79(5). Available at http://www.ecs.soton.ac.uk/ harnad/Papers/Harnad/harnad00.citation.htm.

[7] Hitchcock, S., Carr, L., Hall, W,, Harris, S., Probets, S., Evans, D. & Brailsford, D. (1998). Linking electronic journals: Lessons from the Open Journal project. D-Lib Magazine, 4(12). Available at http://www.dlib.org/dlib/december98/12hitchcock.html.

[8] Ion, P. & Miner, R. (eds) (1999). Mathematical Markup Language (MathML) 1.01 Specification, W3C Recommendation. Available at http://www.w3.org/TR/REC-MathML/

[9] Kofler, M. (2001). MySQL. New York, NY: Springer.

[10] Lagoze, C. & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries, Roanoke, VA. Available at http://www.openarchives.org/documents/oai.pdf.

[11] Lamport, L. & Bibby, D. (1994). LaTeX: A Document Preparation System, 2nd edition. MA: Addison Wesley.

[12] Lawrence, S., Giles, C. L. & Bollacker, K. (1999). Digital Libraries and Autonomous Citation Indexing. IEEE Computer, 32(6), 67-71.

[13] Liu, X., Maly, K., Zubair, M. & Nelson, M. L. (2001). Arc - An OAI service provider for digital library federation. D-Lib Magazine, 7(4). Available at http://www.dlib.org/dlib/april01/liu/04liu.html.

[14] Liu, L., Maly, K., Zubair, M., Hong, Q., Nelson, M., Knudson, F. & Holtkamp, I. (2002). Federated Searching Interface Techniques for Heterogeneous OAI Repositories, Journal of Digital Information, 2(4). Available at http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/.

[15] Maly, K., Zubair, M. & Liu, X. (2001). Kepler - An OAI Data/Service Provider for the Individual. D-Lib Magazine, 7(4). Available at http://www.dlib.org/dlib/april01/maly/04maly.html.

[16] Maly K., Zubair M., Anan H., Tan D., & Zhang Y. (2000) "Scalable Digital Libraries based on NCSTRL/DIENST", Proceedings of ECDL 2000, Lisbon Portugal, pp. 168-180.

[17] Moss, K. (1999). Java Servlets (Second Edition). Boston, MA: McGraw-Hill Companies, Inc.

[18] Oracle. (2001). Oracle InterMedia Server.

[19] Powell, A. & Apps, A. (2001). Encoding OpenURLs in Dublin Core metadata, Ariadne Magazine, Issue 27, Available at http://www.ariadne.ac.uk/issue27/metadata/.

[20] Reese, G. (2000). Database programming with JDBC and Java. Sebastopol, CA: O'Reilly & Associates.

[21] Van de Sompel, H. & Beit-Arie, O. (2001). Open Linking in the Scholarly Information Environment Using the OpenURL Framework. D-Lib Magazine, 7(3). Available at http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html.

[22] Weibel, S., Kunze, J., Lagoze, C. & Wolfe, M. (1998). Dublin Core metadata for resource discovery. Internet RFC-2413. Available at ftp://ftp.isi.edu/in-notes/rfc2413.txt.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 35-43*


# Linking Collection Management Policy to Metadata for Preservation – a Guidance Model to Define Metadata Description Levels in Digital Archives

Maria Luisa Calanag, Koichi Tabata, Shigeo Sugimoto
University of Library and Information Science
Kasuga 1-2, Tsukuba City, Japan
{calanag, tabata, sugimoto} @ ulis.ac.jp

## Abstract

*In an environment of rapid technological change, collection managers face the challenge of ensuring that valuable resources remain accessible when there are changes to the technological context in which those resources are embedded. In this context of requiring "accessibility over time", digital preservation initiatives also demand for interoperability, or as what Hedstrom calls temporal interoperability. But first, libraries, especially in the academic world, need some general guidelines to assist in selectively choosing digital resources which are of great need to collect and preserve. This paper attempts to provide some structure for the concepts and ideas on a general collection management decision guide in the form of a requirements analysis framework that may assist in determining the metadata granularity required for digital resource management within an archive. The objective is for metadata and mechanisms to be shared among digital archives, but policies can be tailored to the requirements of the organization.*
**Keywords:** *digital archives, preservation metadata.*

## 1. Introduction - Interoperability over time

*We need scalable tools and standards for interoperability between archives.*
Margaret Hedstrom

Paul Miller of UKOLN's Interoperability Focus, defines "interoperability" as follows: "to be interoperable, one should actively be engaged in the ongoing process of ensuring that the systems, procedures and culture of an organization are managed in such a way as to maximize opportunities for exchange and re-use of information, whether internally or externally".

### 1.1 Layers of interoperability

To achieve interoperability, the most practical way is to comply with standards. However, implementers often have to choose between standards and how to apply these high-level principles and standards to the "real world". From a "layer" model view, technical interoperability might be seen as the base on which other layers are built, where XML is seen as the standard facilitating technical interoperability. On the other hand, initiatives such as the Dublin Core Metadata Initiative (DCMI) and the Resource Description Framework (RDF) are seen as facilitating semantic interoperability (Johnston, 2001).

Hedstrom (2001) describes the concept of temporal interoperability as the ability of current systems or legacy systems to interoperate with future systems that may use new formats, data models, languages, communication protocols, and hardware. Temporal interoperability promises to make the digital archives of the future as interoperable as today's digital libraries.

Johnston (2001) further mentioned that there is also the aspect of inter-community interoperability that has to be considered, and that "collection description" could be a mechanism to attain this type of interoperability. Libraries have traditionally concentrated on the description of the individual items of their holdings in the form of catalog records. In a networked environment, there is a growing recognition of the value of complementing this item-level description with descriptions of higher-level aggregates of material. Collection descriptions can give an overview of groups of otherwise "uncataloged" items. Managers of archival and museum resources have traditionally made greater use of description at higher levels. As one example, the RSLP Collection Description project

developed a collection description metadata schema which was based in part on the Dublin Core metadata element set, and an RDF implementation of that schema. The RSLP schema can be used in our proposed model. Sections 2.3 and 2.4 discuss more about collection-level descriptions.

**1.2 Existing standards**

Preservation metadata is comprised mainly of what most people would refer to as descriptive, administrative and structural metadata. There are a huge number of metadata initiatives, and it is difficult to figure out how these initiatives can work together. Dale (2002) explained how initiatives like the Open Archival Information System (OAIS), the Open Archives Initiative (OAI), the Online Information eXchange (ONIX), and the OpenURL could potentially fit and work together in the following ways. OAIS provides a general framework and an information model, with the U.K.'s Cedars project leading the way in developing an OAIS-based metadata specification. The OAI standards, on the other hand, defined ways in which descriptive metadata (Dublin Core) could be shared between organizations. ONIX is a standard for the representation and communication of product information from the book industry. OpenURL is a standardized format for transporting bibliographic-type metadata between information services and could be used as a basis for reference linking. It is possible that an OpenURL could link to an OAIS Dissemination Information Package (DIP). As for a 'wrapper' that would be able to link content and all this metadata together, there is now the XML-based Metadata Encoding and Transmission Standard (METS) initiative, now coordinated by the Research Libraries Group (RLG). METS is one option for encoding all of the information that would make up OAIS Information Packages. METS objects and OAIS Archival Information Packages (AIPs) would contain inside them all of the other types of metadata.

The METS schema builds upon the work of The Making of America II project (MOA2) and provides an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories. A METS document consists of four main sections: Descriptive metadata, Administrative metadata, File groups, and Structural map. The European Commission co-funded Metadata Engine (METAe) Project , for instance, decided to adopt the METS schema as its standard output schema for several reasons. Firstly, METS emerged from the MOA 2 project, hence, it has a strong practical implementation aspect. Second, it has an open and flexible structure. Third, it is publicly available at the Library of Congress, and most of all, it is a well-described schema.

**1.3 Scope and some challenges of web preservation**

Since digital libraries are dynamic and widespread, with content, structure, location, delivery systems, and users changing frequently and instantaneously, they require new thinking and models for information management, access, use, and long-term archiving and preservation (Griffin, 2000).

Traditional libraries stress:

- Service
- Selection, organization, structure for access
- Centralization, standards
- Physical objects and standard genres

Contemporary technological capabilities (e.g. WWW) stress:

- Flexibility, openness
- Rapid evolution
- Decentralization (geographic, administrative)
- Digital objects, old and new genres

Digital preservation and digital archiving have been used interchangeably. Both terms mean taking steps to ensure the longevity of electronic documents. The 1996 Task Force Report on Archiving of Digital Information produced by the Commission on Preservation and Access (now the Council on Library and Information Resources) and the Research Libraries Group (RLG) considers long-term preservation as similar to archiving, and actually identifies digital archives, rather than digital libraries, as the unit of activity for the long-term preservation of digital materials. How does a library differ then from an archive? In the traditional sense of the word, these two institutions are usually distinct and separate entities with libraries focusing on the access function, and archives concerned with preservation. In the networked environment though, it would seem that archives are considered worthless without an access functionality or service.

The Internet Archive, for instance, started out simply as an "archive" according to the definition above. It attempted to collect all publicly accessible web pages, and these were "dumped" into a computer system with no organization or indexing. Even then, the fact is that without the vision of Brewster Kahle and his project's automated approach, these web materials would already have been lost. The nice thing is that an "Internet Library" service has been launched by the Internet Archive in 2001 through its Wayback Machine which now allows people to access archived versions of web sites, although it is still not a perfect system.

To be able to preserve web publications, it is necessary to know the construction of the web and some definitions used to describe it. The web is a way of viewing pieces of information located in different places on the Internet as if they were one large indexed document by using hypertext and multime-

dia technique. This means that in a way it is impossible to preserve single publications completely because they have links pointing to other documents, which in turn link to others. Long-term preservation of the web seems to be hard to achieve, since a web page could not be preserved on paper or microfilm because the hypertext and multimedia techniques embedded will get lost and can never be retrieved again. Hence, the authors are also interested and concurrently looking into some ways on how to maintain this link functionality "over time" with the use of metadata.

### 1.4 A requirements analysis framework for formulating metadata guidelines for collection management & preservation

Collection management policies that deal with digital materials present one of the most critical challenges collection managers have to face. This will not be limited to technical issues only, but equally important are the organizational and management issues. Preservation decisions must be done at an early stage of the lifecycle of resources, since delays in taking preservation decisions can later result in preservation requirements that are more complex and labor intensive. Therefore, there is a strong need to establish guidelines that can assist collection managers in recording the appropriate level of metadata for collection management and preservation. The goal of this paper is to offer a requirements analysis framework which associates collection management policy with metadata to help collection managers define appropriate metadata based on their own requirements. The desired effect is for collection-level metadata and mechanisms to be shared among digital archives, but policies can be tailored to the requirements of the organization.

## 2. Collection management&preservation

*"The next great revolution in libraries will be in collection development."*
Stanley Chodorow

"Collection management policy" is a broader term than collection development, intended also to include storage, maintenance, de-selection and preservation. It is an important tool for defining what materials are of long-term interest to the collection. It needs to specify the acceptable level of functionality that has to be preserved if a digital object is to be retained. Such decisions will influence the level and method of access that will be necessary for the object as well as the level of preservation metadata required for long-term retention. For digital materials, value judgments made by the archivist/collection manager will determine what level of functionality needs to be retained. The Cedars Project has coined the term "significant properties" to describe those components of a digital object deemed necessary for its long-term preservation. Determining the significant properties of a digital object, i.e. the acceptable level of functionality, will dictate the amount of information or "metadata" that must be stored alongside the bytestream (the Data Object) to ensure that it remains "renderable" over time. How much specificity can be added to the metadata description, while maintaining broad applicability at the same time – is the authors' motivation in proposing the use of metadata description levels in this paper.

### 2.1 The responsibility for web preservation

Digital preservation is defined as the managed activities necessary for ensuring the long-term maintenance and continued accessibility of digital materials. It involves two main functions: the long-term maintenance of a bytestream and continued accessibility to its contents. Effective lifecycle management depends on a proactive approach and the cooperation of a number of stakeholders including content creators (See Figure 1 for the lifecycle diagram).

According to Allen (2000), "The management of digital collections is becoming a core Web-based service". He used the acronym SOAP to describe his essential elements in collection management, which are Selection, Organization, Access, and Persistence. It is also realistic to assume that we can not depend on data creators to preserve their own work because they lack both the power and the motivation to do so. Casey (1998) points out that the creator is rarely the "owner" of the server space where a site is located:

*More often than not, Web site stability relies on who "owns" the server space where a site resides. Ownership means that the author of the Web site has control over use of the space as long as the content is within the policies of the administration of the server… Many folks in the academic world use the space allowed them on their university accounts to post Web pages. They cannot claim ownership of this space, just the right to borrow it for as long as they are associated with the institution or according to the Internet usage policy of the university. The irony is that many of these sites possess the content and quality that librarians want to preserve.*

Preliminary results of a survey (Greenstein et al., 2001) issued by the Digital Libraries Federation (DLF) to its members discussed the library's relative role in creating, providing access to, and preserving digital assets within the university that contribute new forms of scholarly communication (e.g. e-journals, e-print repositories, digitized content, etc.). Many units within the university are taking responsibility for the production of digital content that contribute new forms of scholarly communications. The

library is primarily responsible for the production of that content based on library holdings. Responsibility for other such content is widely spread across units with academic departments taking primarily responsibility for e-print repositories, e-journals, and distance learning materials. IT and academic computing departments have limited responsibility for production of digital information content of any kind.

The library though has a greater role in providing access to this content much more than creation of content. It is primarily responsible for providing access to digitized library content, to e-journal content, to e-books and to e-prints. Where preservation of such content is concerned, only the digitized library holdings appear at all to be secure. Most respondents to the DLF survey claim that the library takes responsibility for the preservation of these holdings, but other kinds of digital content such as e-journals and e-prints are apparently at risk.

### 2.2 Lifecycle management of digital materials

In traditional records management, the term 'information lifecycle' has long been used to describe the processes related to the creation and management of information. This concept is illustrated in Figure 1 (Brown, 2000). Preservation of digital materials needs to be an integral part of digital collection management and must therefore be incorporated into the overall management of an organization from acquisition through to preservation. It requires active management that begins at the creation of the material and depends on a proactive approach by digital repositories and the cooperation of stakeholders.

### 2.3 Collection descriptions

In the library domain, discussion has tended to focus on so-called "item" level metadata (i.e., descriptions of individual books, articles, and so on). The new environment brings about new requirements. The broker needs to have access to various types of metadata to support its operation. This is data about its environment and the resources in it. It should be clear that metadata is of central importance in distributed information environments.

Typically information objects exist in collections, where a collection comprises similar information objects. These collections might be databases, websites, document supply centers or libraries. They may be particular collections within a library, or the catalog for such collections. Such collections are also, of course, information objects, and collections may contain other collections. Collections will also have different terms and conditions associated with their use. Typically collections will be managed by organiza-

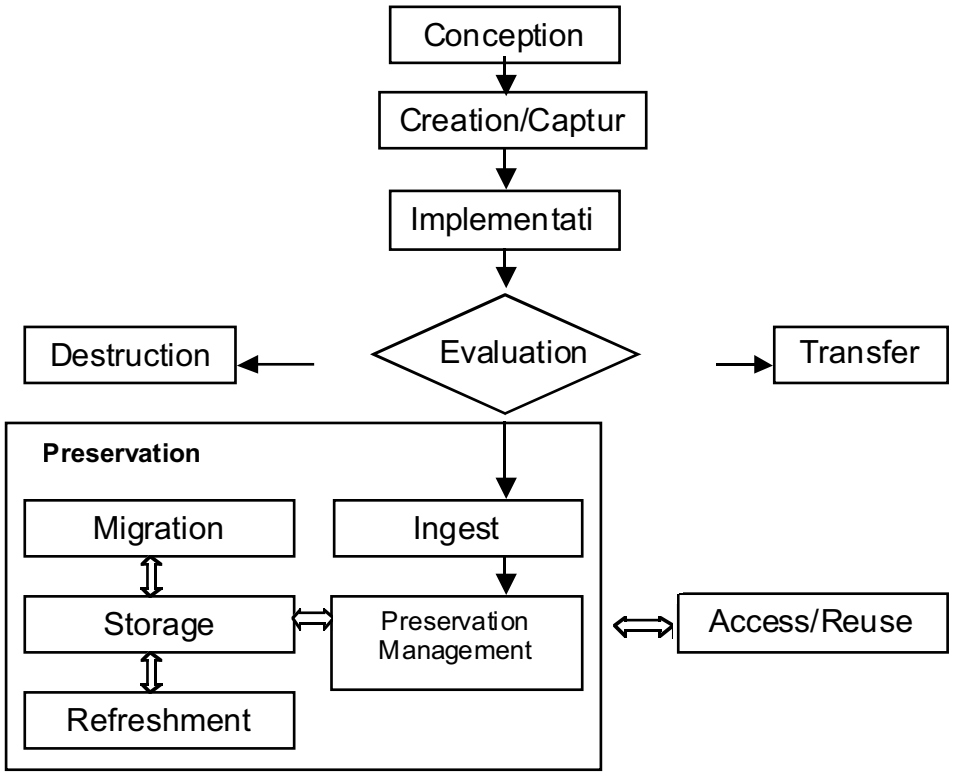



**Figure 1. The Information Lifecycle (Used with permission from the Cedars Project. Cedars Guide to Digital Collection Management, 2002)**

tions. Information objects may be data or metadata.

Those who used Dublin Core (DC) expressed a strong need for item-level access, and somewhat less concern for grouping items into definable collections or sub-collections. There seemed also to be more uniformity of size and type among their materials. DC is cheaper to work in than MARC because of its limited element set. Those who used the Encoded Archival Description (EAD) standard wanted to organize items into collections and sub-collections, and saw the items just like chapters in a book or articles in a journal. A lack of uniformity of size and type within a collection also made EAD attractive. EAD is also cheaper to work in than MARC, because large numbers of items can be grouped (Seadle, 2001).

Archival description is the equivalent in archivology to cataloging in librarianship. There are important differences of principle and practice between these two fields. The definition from the General International Standard Archival Description (ISAD(G), 2001) makes use of two important concepts underlying archival management, which are as follows:

- The principle of representation: Because original archival materials cannot be organized for direct physical access by users, they have to be managed and retrieved by using representations. These have to contain the right data to allow for their effective use in the various management functions.
- The unit of description: The basic unit in archival management is taken to be the group ("fonds" in international usage, and also often called a "collection"). Most often, a group is a large body of materials that can be subdivided into subordinate entities. It would be normal, therefore, for an archive group to have a description representing the whole group, followed by a number of interlinked descriptions of its components. Generally, archival descriptions must contain information on the provenance, background, and context of the materials. It is, in principle, not possible to describe archival materials in terms of their contents and physical form alone. Provenance information includes a history of the administration or activity that caused the archives to be created and explains how they were used during the period when they were current records.

### 2.4 Linking policy to metadata

By merging traditional collection levels (Table 1) and collection level descriptions for digital resources which we call "Persistence levels" (Table 2) in the form of a matrix (Table 3), this can serve as a good starting point for developing a method of linking policy to metadata (Calanag et al., 2001). In addition, a set of values can be chosen for each combination according to the degree to which digital materials are persistent based on LeFurgy's (2002) definitions. Persistence is based on consistent and transparent rules for description and structure, standardized file formats, and so forth. In general terms, LeFurgy said that degrees of persistence can be represented in three categories (LeFurgy, 2002). In Table 3, these confidence ratings are what we considered as "Preservation requirement levels" in this paper.

- High (H): Fully persistent materials that enable high confidence for ongoing preservation and access.
- Medium (M): Partially persistent materials that enable medium confidence for ongoing preservation and access.
- Low (L): Marginally persistent materials that enable low confidence for ongoing preservation and access.

Given that persistence is closely tied to the clarity and consistency with standards by digital resources, it follows that materials that are highly structured tend to be inherently easier to preserve and access over time. Conversely, less structured materials tend to be harder to manage. In addition, persistence can also be tied to resource availability in terms of the digital object's persistent identifier.

The authors propose that these three Preservation requirement levels (High/Medium/Low) may determine the granularity of the preservation metadata that will be required to ensure that the digital materials will be preserved and accessed over time. In other words, a choice among High/Medium/Low can be associated with item-level, class-level, or collection-level preservation metadata, respectively (see Table 4). As shown in a Sample Policy Table (Table 3), a general rule of thumb is that we go from High to Low as the persistence levels gain lower confidence and stability. Collection manager-defined default ratings or a blank space(s) denoting Not Applicable can be assigned according to the institution's policy.

## 3. Digital archives in academia

*"Universities are becoming publishers and they need to take responsibility for their own output."*
Cedars Final Workshop summary

There is some need for institutional responsibility from universities, especially with regard to local scholarly material, learning objects and institutional records. Cedars, for example, had focused on the incoming digital acquisitions of research libraries and the intellectual content created by institutions, both digitized and "born digital". Preservation was about the continued accessibility of the content of digital resources, and was focused on the content rather than any particular medium. One major difference between traditional collection management strategies and that needed for digital information is that consideration of preservation requirements needed to hap-

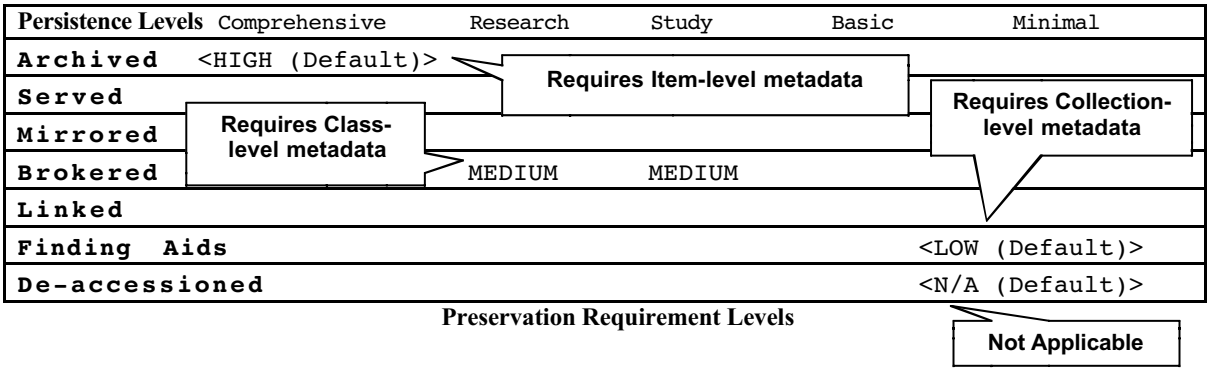
**Table 1. Collection levels**

| Levels | Description |
|---|---|
| Comprehensive | A collection to include all significant works of recorded knowledge in all applicable languages for a defined and limited field. |
| Research | A collection which includes the major dissertations and independent research, including materials containing research reporting new findings, scientific experimental results, and other information useful to research. |
| Study | A collection which is adequate to support undergraduate and most graduate course work, and to maintain knowledge of a subject required for limited or general purposes. |
| Basic | A highly selective collection which serves to introduce and define the information available elsewhere. |
| Minimal | A collection in which few selections are made beyond very specific works. |

**Table 2. Persistence levels**

| Levels | Description |
|---|---|
| Archived | Material is hosted in the library, and it intends to keep intellectual content of material available permanently. |
| Served | Material is hosted in the library, but no commitment to keeping it available. |
| Mirrored | Copy of material residing elsewhere is hosted in the library, and it makes no commitment to archiving. Another institution has primary responsibility for content and maintenance. |
| Brokered | Material is physically hosted elsewhere and maintained by another institution, but the library has negotiated access to it; includes metadata and links in the catalog, and library users can locate and cross-search it. |
| Linked | Material is hosted elsewhere, and the library points to it at that location; no control over the material. |
| Finding Aids | Electronic finding aids and metadata held by the library to facilitate discovery and searching; this metadata is associated with the library's digital collections or elsewhere, but may be stored, managed and maintained separately from them. |
| De-accessioned | Accessioned resources that have not been retained after review. |

**Table 3. Putting it all together:**
**A Requirements analysis matrix linking policy and metadata – A Sample Policy Table**




In using this matrix, a general rule of thumb is that we go from High to Low as the persistence levels gain lower confidence and stability. Collection manager - defined default ratings or Not Applicable <N/A>ratings can be assigned according to the institution's policy.

pen much earlier in a resource's life cycle. Decisions taken at each stage in the lifecycle would influence options at other stages. It follows, therefore, that creators play a significant role in digital preservation.

The most likely collection model would be distributed, but there would be a need for transparency as

to which organizations are preserving what materials and clarification of roles and responsibilities. We would have to adapt to high volumes of information which would stress the importance of distributed solutions and the automation of ingest and metadata capture processes. There would also be a need to find and manage the information, based on metadata and persistent identification. Another major challenge in the academic sector would be e-prints and e-theses. The scale of the challenges faced would mean future archiving would be distributed.

### 3.1 Persistent archive architecture

Archivists rely on persistent archives to support all aspects of data collection management. Persistent archives provide a mechanism needed to support distributed data access across heterogeneous data resources (Moore, 2002; Ludascher, et.al., 2001). Using concepts and terminology from the Open Archival Information System (OAIS) reference model, Figure 2 shows a digital archive architecture that can be built around XML-based standards and technologies.

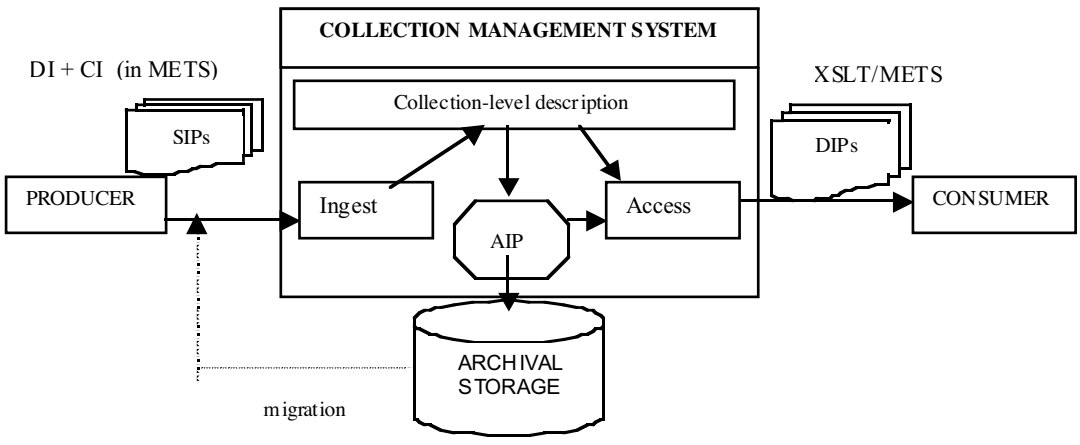
First, the producer and the archive need to agree on the submission policies (e.g., acceptable submission formats, specifications on what are to be preserved, access functions, and other legal requirements), and the preservation policies. General preservation decisions can be made based on the matrix presented in Table 3 which will serve as a requirements analysis framework. Then, the producer can ingest these SIPs (METS-encoded Submission Information Packages = Descriptive Information + Content Information) into the Collection Management System where they are assigned the appropriate metadata at the granularity level based on the requirements analysis framework. The "highly persistent" (H) the resource is considered to be, the more detailed preservation metadata should be to allow the resource to be emulated, for example, on future platforms. Once these additional information are added to Descriptive metadata, SIPs are transformed into AIPs (Archival Information Packages) which are put into archival storage. Migration of AIPs in the archive is simply a regular refreshing process (for now) to prevent obsolescence of AIPs. The bitstream of the content will remain unchanged. Only new provenance metadata will be added every time medium migration is done.

The levels of metadata granularity are described in Table 4 which shows their equivalence to the preservation requirement levels.

### 3.2 Preservation metadata at the three granularity levels

Figure 3 presents a simple collection description model to provide a view of the framework into which the metadata granularity level fits. Most of the preservation metadata elements enumerated in Appendix 1 have been recommended by the OCLC/RLG Working Group on Preservation Metadata (2002) in their latest report. Grouping the metadata elements according to the three granularity levels, is one possible categorization proposed by the authors.

This is how the proposed "collection management decision guide" (Table 3) can be applied. Default ratings can be set for certain combinations. However, let us take a specific example, a HIGH rating has been assigned to the combination SERVED + STUDY by the collection manager. This means that Item-level description or metadata should be provided for each




Legend:
DI - Descriptive Information
CI – Content Information
SIPs – Submission Information Package(s)
AIP – Archival Information Package
DIPs – Dissemination Information Package(s)

**Figure 2. Digital archive architecture**

**Table 4. Mapping between Preservation requirement levels and metadata granularity**

| | Metadata granularity | Description |
|---|---|---|
| High | item-level metadata | Individual digital objects are packaged into the Content Information (CI). |
| Medium | class-level metadata | Structural information is handled; this metadata describes types of object attributes, and aggregation information (Context Information) |
| Low | collection-level metadata | Can be added to the Descriptive Information (DI) and in this paper, this also refers to the RSLP collection description schema |

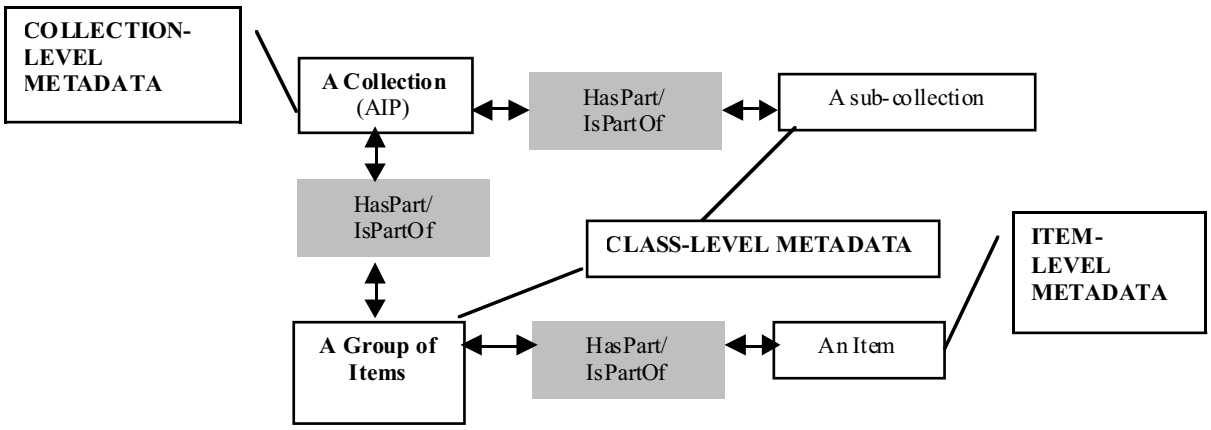



**Figure 3. Collection description model**

item in the aggregation or set. To designate a HIGH rating entails a big responsibility and commitment on part of the institution since very detailed metadata has to be provided or generated (see Appendix 1). Decisions have to be given much thought by the collection manager, and one main criterion that can guide decision-making is the persistence of materials. On the other hand, if it has been decided that LINKED + BASIC =LOW, then it means that Collection-level description or metadata should be used. These collection-level descriptions or metadata can then be shared among digital archives for cross-searching, access and re-use.

## 4. Conclusions

The authors have laid down a collection management guide in the form of a requirements analysis matrix for general applicability in the academic environment, where preservation policy decisions can be made according to local requirements. It also prescribed a digital archive architecture that can be used in distributed environments which can serve as a mechanism for institutions to coordinate their digital preservation activities while at the same time, retaining the flexibility to meet their local needs.

In selecting materials for preservation, evaluation decisions might reflect technical issues including the quality of the data object and its existing metadata, and the technical environment, both hardware and software, needed to access and use the data object. According to the persistence of resources as determined by collection managers and/or information producers, this paper prescribed a way to ensure that documentation will be preserved so that environments can be reconstructed for future "processability" or accessibility.

For organizations taking responsibility for the long-term preservation of digital materials, a written and up to date collection management policy is critical. It provides an important tool for the collection managerby inviting consideration of all the relevant issues early in the lifecycle of digital materials within their scope.

Two vital criteria for preservation are to ensure that the preserved digital object can be found, and that the preserved digital object can be understood. For these criteria to be met, it is vital that each preserved digital object has a unique and persistent identifier. For their future work, the authors are currently conceptualizing a mechanism for encoding preservation metadata in a URL that offers context-sensitive links that should lead to the appropriate versions of the resource that the user needs.

## References

Allen, Robert B., 2000. Digital libraries - a cornerstone of web-based services. *In*: International Conference of Asian Digital Library 2000, Seoul.

Brown, Adrian, 2000. Digital Archiving Strategy. English Heritage. http://www.english-heritage.org.uk/default.asp?wci=WebItem&WCE=544

Calanag, Maria Luisa, Sugimoto, Shigeo, and Tabata, Koichi. A metadata approach to digital preservation. *In*: *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001. Tokyo: National Institute of Informatics, October 2001*.

Casey, Carol, 1998. The Cyberarchive: a look at the storage and preservation of web sites. *College and Research Libraries* , 59 (4), 304-310.

Chodorow , Stanley, 2001. Scholarship, Information, and Libraries in the Electronic Age. *In*: Marcum, D.B., ed. *Development of digital libraries: an American perspective*. Westport, Conn.: Greenwood Press.

Dale. Robin, 2002. Cedars final workshop. Manchester: UMIST.

Day, Michael and Jones, Maggie, 2002. Cedars (CURL Exemplars in Digital Archives) final workshop: executive summary. Manchester: UMIST.

Greenstein, D., Thorin, S. and McKinney, D., 2001. Draft report of a meeting held on 10 April in Washington Dc to discuss preliminary results of a survey issued by the DLF to its members.

Griffin, Stephen M., 2000. The Digital Libraries Initiative: a USA federal program of research and applications. *In: Digital Libraries.* No. 18.

Harrod's Librarians Glossary, 1999. Compiled by Ray Prytherch. Gower.

Hedstrom, Margaret, 2001. Exploring the concept of temporal interoperability as a framework for digital preservation. University of Michigan.

ISAD (G): General International Standard Archival Description. Ottawa: International Council on Archives, 2000.

Johnston, Pete, 2001. Interoperability: supporting effective access to information resources. *Library & Information Briefings*. London: South Bank University.

LeFurgy, William G, 2002. Levels of service for digital repositories. *D-Lib Magazine*.

Ludascher, B., Marciano, R., and Moore, R., 2001. Preservation of digital data with self-validating, self-instantiating knowledge-based archives. *SIGMOD Record*, vol. 30, no. 3, pp. 54-63.

The Making of America II Testbed Project. http://sunsite.berkeley.edu/MOA2/moaproposal.html

Metadata Encoding and Transmission Standard (METS): an overview and tutorial. http://www.loc.gov/standards/mets/METSOverview.html

Moore, Reagan W., 2002. Persistent Archive Concept Paper Draft 2.0. San Diego Supercomputer Center: Persistent Archive Research Group Global Grid Forum.

The OCLC/RLG Working Group on Preservation Metadata, 2002. Preservation metadata and the OAIS information model: a metadata framework to support the preservation of digital objects. Ohio: OCLC Online Computer Library, Inc.

The Prism project: vision and focus. Prism working paper 2000-02-07 http://prism.cornell.edu/Publications/WorkingPapers/Visions.htm

Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems. May 1999. http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

Seadle, Michael, 2001.Sound practice: a report of the best practices for digital sound meeting, 16 January 2001 at the Library of Congress. *RLG DigiNews*, April 15, 2001, vol. 5 no.2 http://www.rlg.org/preserv/diginews/diginews5-2.html

Task Force on Archiving of Digital Information. Preserving Digital Information: Report of the Task Force on Archiving of Digital Information. 1996.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 45-52*


# Semantic Web Construction: An Inquiry of Authors' Views on Collaborative Metadata Generation

Jane Greenberg
School of Information and Library Sciences
University of North Carolina at Chapel Hill
janeg@ils.unc.edu

W. Davenport Robertson
National Institute of Environmental Health Sciences
robert11@niehs.nih.gov

## Abstract

*Increasing the amount and quality of metadata is essential for realizing the Semantic Web. The research reported on in this article addresses this topic by investigating how resource authors might best collaborate with metadata experts to expedite and improve metadata production. Resource authors, working as scientists at the National Institute of Environmental Health Sciences (NIEHS), were surveyed about collaborating with metadata experts (catalogers) during the metadata creation process. The majority of authors surveyed recognized cataloger expertise is important for organizing and indexing web resources and support the development of a collaborative metadata production operation. Authors discovered that, as creators of web resource intellectual content, they too have knowledge valuable for cataloging. This paper presents the study's framework and results, and discusses the value of collaborative metadata generation for realizing the Semantic Web.*
**Keywords:** *Semantic Web, Collaborative Metadata Generation, Human generated metadata, Dublin Core, National Institute of Environmental Health Sciences (NIEHS), Government Agencies.*

## 1. Introduction

Envisioned by Tim Berners-Lee, the inventor of the World Wide Web (web), and further defined by a number of key researchers and visionaries, the Semantic Web aims to bring structure to the web's meaningful content. The goal, as Lassila et al. (2001) explain, is to build a structured environment where software agents roam and carry out sophisticated tasks, such as arranging *all* the components of a summer vacation, from air travel and hotel to a night on the town. Structured knowledge representation underlying the Semantic Web needs to be built upon trusted metadata—that is accurate, consistent, sufficient, and thus reliable metadata.

Although researchers agree creating trusted metadata is fundamental to realizing the Semantic Web, examining partnerships among persons involved in metadata creation does appear to be a major focus. A probable reason for this predicament is the need to *first* clarify the Semantic Web's conceptual design, an undertaking being documented via numerous theoretical and practical discussions (see links from: (www.w3.org/2001/sw/). Another possible factor is the need to invent and test Semantic Web languages, or what may be thought of as infrastructure technologies (e.g., Resource Description Framework (RDF) (www.w3.org/RDF/), DAML (DARPA Agent Metadata Language) + OIL (Ontology Inference Layer) Reference Description (www.w3.org/TR/daml +oil-reference), and now OWL (Ontology Working Group Language) (Patel-Schneider, 2002). This second focus is evident by research presented at the recent Semantic Web Conference (ISWC2002), Sardinia, Italia (see conference proceedings at: link.springer.de/link/service/series/0558/tocs/t2342.htm). These research emphases are critical to the Semantic Web's development, although they do not specifically address the fact that a vast majority of web content is not semantically encoded with the metadata required for agent roaming and automatic processing activities.

If the amount and quality of web content metadata is to increase, Semantic Web efforts need to also prioritize metadata generation research. Important foundation work designing metadata schemas (e.g., Dempsey et al. 1997) and developing metadata tools

(For example, see Dublin Core tools at: www.dublincore.org/tools/) is in progress. Paramount now is the need to discover the best means for efficiently producing good quality metadata, drawing from both human and automatic processes.

## 2. Human-Metadata Generation

*Human-metadata generation*, the focus of this study, takes place when a person is responsible for the identification and assignment or recording of resource metadata. Human-metadata generation is often explained by distinguishing it from *automatic-metadata generation*. In the first case a person intellectually manages in the metadata generation, whereas in the latter case a machine-based algorithm automatically extracts metadata from the resource content. Both methods have strengths and weaknesses, and experts, particularly in the area of subject indexing, agree that the most effective results can be achieved through methodological integration (e.g., Schwartz 2000, p. 152). Although empirical evidence is limited, it appears that the variety of persons involved in metadata generation also exhibit different strengths and weaknesses, making it likely that the best results will be achieved through skill integration.

This research focuses on potential collaboration between metadata experts (here after referred to as experts) and resource authors (hereafter referred to as authors). These classes of persons have been selected as they are among two of the most active producers of descriptive metadata. A discussion of the persons involved in metadata generation is found in Greenberg (2002). Descriptive metadata includes elements, such as "title", "author/contributor", and "subject"; these elements provide surrogates for information resources and facilitate discovery. *Information resources* are objects housed in digital and physical libraries, museums, archives, and like information centers.

Experts include catalogers, indexers, and other persons having formal education, earning an advanced degree in information or library science. They are often preferred metadata creators because their ability "to make sophisticated interpretative metadata decisions and work with classificatory systems" (Greenberg 2002) aids in the production of high quality metadata (Weinheimer 2000). Experts' skills are, however, insufficient when addressing common web resource problems stemming from the absence of "title pages" and other standard bibliographic features, which are heavily relied on in cataloging. As third-party metadata creators, experts may not be privy to resource details needed for creating descriptive metadata.

Authors include persons who produced the intellectual content of the resource being cataloged. They are intimate with their creations and have knowledge of *unrecorded* information valuable for producing descriptive metadata. An example is "date of creation" metadata. A scientist/author may know when a report was originally published, although the web version may not show this information. Exploratory research demonstrates to some degree that authors can produce acceptable metadata (Barrueco & Krichel 2000, Greenberg et al. 2001). Further evidence is found in that commercial information databases (e.g., *Dissertation Abstracts*) index resources with abstracts, keywords, and other author-generated metadata. In fact, many publishers of scientific journals require authors to submit subject "keywords" with their manuscripts. A limitation with authors as metadata creators is that they may lack knowledge of indexing and cataloging principles and practices, and are more likely to generate insufficient and poor quality metadata that may hamper resource discovery (Milstead & Feldman 1999, Thomas & Griffin 1999, Weinheimer 2000).

The field of information and library science has a substantial body of research studying automatic and human indexing, a recent summary of which is found in Anderson and Pérez-Carballo (2001). Additionally, metadata generation tools experiment with the integration of automatic and human processes e.g., www.lub.lu.se/tk/metadata/dctoollist.html. In efforts to realize the Semantic Web, it makes sense to further extend comparisons and integration activities to collaboration among different classes of persons generating metadata—the goal of this research.

## 3. Towards A Collaborative Metadata Generation Framework

Collaborative metadata generation, as defined in this study, is the *joint production of web resource metadata*. While official collaborative metadata generation programs appear scarce, collaboration research, together with long-standing indexing practices and recent web-based initiatives, provide a framework for developing such an operation. Several of these developments are highlighted below:

***"Collaboration" research.*** A growing area of research focuses on collaboration between "system designers" and "potential users" during information system design activities (e.g., Sonnenwald and Lievrouw 1996). This work examines social and behavioral issues that arise when "technical experts" (system designers) and "clients" (persons for whom a system is being designed and who have intimate discipline knowledge) collaborate. Results provide insight into issues that may arise when experts (e.g., catalogers) and authors, who are domain experts with respect to their creations, collaborate during metadata creation.

***De-facto collaborative metadata generation.*** As highlighted before, scientists and scholars generated "abstracts", "keywords" and other metadata for their

publications. Commercial databases adopt and enhance this metadata for access. Frequently a metadata expert conducts authority control work to standardize subject and name-headings. This framework is one of *economy*, allowing metadata experts (generally indexers) to take advantage of author knowledge and devote their valuable and more costly time to metadata activities requiring professional training. The partnership may be viewed as a *de-facto* collaboration rather than an *active* collaboration because of the absence of real time communication between author and professional.

***Dublin Core metadata.*** The Dublin Core Metadata Initiative (DMCI) has facilitated the development and use of the Dublin Core Metadata Element Set (1997), a schema comprised of 15 elements deemed essential for resource discovery (Weibel 1995, Duval et al. 2002). An underlying principle is that this schema is simple enough for nonprofessional use. Authors can create metadata working with simple templates or editors, and experts can subsequently enhance this metadata following a more complex schema or by performing authority control work. OCLC's Cooperative Online Resource Catalog (CORC) (www.oclc.org/corc) project provides framework for this type of collaboration.

***Open Archives Initiative.*** The Open Archives Initiative (OAI) (www.openarchives.org/) promotes interoperability standards that facilitate efficient access to web content and other forms of electronic documentation. The OAI has adopted the Dublin Core metadata schema. OAI projects use a variety of metadata generation techniques, including metadata produced by experts or authors. Metadata from any OAI compliant initiative can be harvested and placed in *single service*, and may result in a collection of metadata generated by experts for some resources and by authors for other resources. Integrating expert and author produced metadata records, post-metadata creation, may be viewed as a partnership in that both parties (experts and authors) are contributing to a larger pool of resource representation—generally for a particular domain. It's likely that some OAI projects carry out collaborative metadata generation during the initial metadata production stage, although documentation is limited.

***Metadata tutorials.*** Metadata initiatives associated with the web expand well beyond the traditional library environment to other information communities (e.g., commerce, health science, and geo-science). As part of this development, experts have been called upon to write schema specifications and design tutorials instructing authors and other persons about metadata creation. Additionally, many HTML guides instruct web developers and resource authors about the creation of meta tags, often highlighting the "keyword," and "description" tag (e.g., Dr Clue's HTML/CGI Guide (http://www.drclue.net/F1.cgi/HTML/META/META.html). Expert designed tutorials providing metadata creation guidance to authors and other non-metadata experts provides a form of collective metadata generation that may have implications for collaborative activities.

The developments reviewed here provided a framework for this paper's examination of authors' attitudes about collaborative metadata generation involving experts.

## 4. Research Goals

This study was conducted to gain insight into authors' perceptions about collaborative metadata generation. The study was conducted as part of a larger ongoing study that is examining human and automatic metadata generation methods. An underlying goal of this study is to assist the National Institute of Environmental Health Sciences (NIEHS) metadata project. A broader objective is to share these results with similar organizations aiming to increase the amount and quality of metadata, while contributing to the Semantic Web's construction. Questions guiding the study were:

- Do authors think expert assistance would be useful during the metadata creation process?
- What communication methods do authors prefer in a collaborative metadata generation operation?
- What types of metadata are authors most likely to seek expert help generating in a collaborative environment?

## 5. Method

The survey method was used to gather data on author's views about collaborative metadata generation. This survey was supplemented by data gathered via a participant profile questionnaire and a post-metadata creation questionnaire implemented in a larger ongoing metadata generation study, which will be reported on in a future paper.

The test domain was the National Institute of Environment Health Sciences (NIEHS), an Institute of the National Institutes of Health (NIH), which is a component of the U.S. Department of Health and Human Services (DHHS). Participants were NIEHS scientists who had created Dublin Core metadata records in the larger metadata generation study for at least one of the web resources they had authored. Thirty-four scientists were each sent a printed copy of the collaboration survey and printed copies of the metadata records they had created in the larger study. The metadata records were reproduced on yellow paper to distinguish them from the collaboration survey and remind participants that, approximately three-months earlier, they had produced at least one Dublin Core metadata record for their web resource. The survey materials were sent via NIEHS interdepartmental mail with the assistance of library staff and student interns, who are active members of the

NIEHS metadata team. Printed survey materials were used instead of electronic materials because library staff indicated that this would most likely result in the highest return rate.

The survey was brief and included a series of questions asking participants if they thought cataloger assistance would be useful during metadata generation, if so—through what methods would they prefer to communicate with a cataloger, and for what types of metadata generation might they seek expert help. An open question at the end asked participants to describe other ways they envision scientists collaborating with catalogers to generate metadata. Participation in the study was optional.

The survey was efficiently designed on a single page that could be folded in-half upon completion to display the library director's return address. The design made it possible for participants to answer the survey and easily place it in inter-departmental mail without the complications of finding an envelope. A period of two weeks was given for survey completion, during which time the library director sent out two e-mails encouraging scientists to respond.

## 6. Results

Nineteen NIEHS scientists responded to the collaborative metadata generation survey. As indicated under methodology, the collaboration survey was sent to 34 scientists participating in the larger study. Of the 19 responses received, 18 (52.9% of the 34 originally distributed) were useful for data analysis. One returned survey was eliminated from data analysis due to a failure to answer any of the questions.

Results of data analyzed for this study fall into two categories: 1) Participant background and metadata knowledge, and 2) Collaborative metadata generation survey results.

### 6.1 Participant background and metadata knowledge

Participant assessment was based on data gathered via a participant profile questionnaire and the post-metadata creation questionnaire implemented in the larger study noted above. This data was culled to provide contextual information for the current study's data analysis and discussion. Of the 18 participants partaking in the collaborative metadata generation survey, nine (50%) search the web daily, six (33.3%) search weekly, and three (16.7 %) search monthly or less than once a month. These results indicate a fairly good comfort level with public and consumer-oriented web technologies (e.g., search engines and consumer web sites, such as Amazon.com). Participants' understanding of the word "metadata" appeared limited with only four (22.2%) of the 18 indicating they had heard the word metadata prior to the NIEHS metadata research. Three of these participants attempted to define metadata, with one response being accurate, giving the definition of "data about information". Limited metadata knowledge was further evidenced by the fact that only one participant had created web resource metadata prior to participating in the NIEHS metadata generation study, although six (33.3%) participants had experience creating HTML (hypertext markup language) documents.

All 18 participants had created at least one metadata record in the NIEHS metadata generation study. This task was completed by inputting metadata into the NIEHS metadata form, which is a simple template based on the Dublin Core metadata schema. Post-questionnaire data gathered after this activity provided insight into participants' views on the value of metadata and metadata creation. A semantic differential scale, on which "1" indicated "with difficulty" and "5" indicated "easily" gathered participants' opinions about the difficulty of the metadata creation task. The majority of participants indicated that it was an *average* to *easy* task, with 16 (88.9%) selecting a "3"or above. A semantic differential scale where "1" indicated "never" and "5" indicated "always" gathered data about participants' views on the need to create web resources metadata. Fifteen participants (83.3%) selected a "3" or above indicating an *average* to *always* support for web resource metadata. A final question asked participants who should create metadata. A check list included the following: "No one," "Authors," "Departmental heads," "Librarians," "Web Masters," "Secretaries" and "Other." Participants were encouraged to select as many options as they would like. Ten participants (55.6%), selected author, whereas 6 participants (33.3%) selected librarians. The results of the profile and post-metadata generation questionnaires show that authors value metadata and believe they should create metadata for their works.

### 6.2 Collaborative metadata generation survey results

The collaborative metadata generation survey gathered data on authors' views about collaborating with a "cataloger" during metadata generation. Data gathered on establishing a collaborative metadata generation program was fairly, although not unanimously positive with 12 of the 18 participants (66.7%) indicating that assistance from a professional cataloger would have been useful during the metadata creation process. Reasons given noted the ability of catalogers to be consistent and make resources accessible to potential users. A few replies also revealed a slight insecurity among participants in terms of their own ability to produced quality metadata. Examples of responses include the following:

- It [cataloger assistance] would ensure that consistent terms were being used across the various programs/initiatives.

- I can't do it adequately without assistance.
- Professional cataloger will/should be up to date and understand communication via this vehicle and enable a broader audience.
- I'm not sure of the best word to use to 'hit' a particular audience.

Six participants (33.3%) indicated that professional cataloger assistance would not have been useful during the metadata creation process. Only one participant provided a textual response, which was related to the fact that he was cataloging a "project website," not a "publication." The response was, "I'm open to ideas, but I think the only webpage that might fit would be publications." Three participants (16.7%) (two supporting cataloger assistance and one not in favor of cataloger assistance) provided textual responses indicating they were confused by the word "metadata" and its meaning. The NIEHS library staff surmised that the confusion stemmed from the survey's use of the words "metadata" and "metadata record", and several responses suggest participants may have actually equated the word "metadata" with HTML encoding required for a webpage.

Despite the noted confusion, participants were clearly cognizant of the fact that they created a "metadata record". Thirteen participants (81.3%) indicated "yes" they wanted to be notified if their metadata record was to be expanded or modified by a professional cataloger in *any way*, while three participants (18.7%) replied "no", they did not want to be notified. (Percentages, based on the population of 18 participants. are given here and in the remainder of this paper). A cross-tabulation was performed to see if there was a correspondence between participants supporting cataloger collaboration and wanting to be notified about metadata records enhanced or modified by a cataloger. The majority of participants (8 of 12, 66.7%) supporting cataloger collaboration wanted to be notified of changes to their metadata record. Five participants (27.8%) supporting cataloger collaboration did not want to be notified of changes; and two participants (11.1%) not supporting cataloger collaboration, wanted to be notified. Three participants did not answer this question. Results for the correspondence analysis indicate that participants' supporting collaboration were enthusiastic about communicating with catalogers even after they created their initial metadata record. These participants may have had a sense of metadata record ownership and/or a desire to learn more about metadata creation from a cataloger. It's likely that participants who did not want cataloger assistance, but wanted to be notified of any changes, had a sense of metadata record ownership for their work (the metadata record they created in the larger experiment and which was given to them on yellow paper for this study). More research is needed to verify these observations, and well as why participants supporting collaboration did not want to be notified of cataloger changes.

Participants were asked about preferred method of communication in a collaborative metadata generation operation. Among options given were: "Face to Face (a personal contact with a cataloger), "E-mail", "Phone", "Web-form" and "Other". Seventeen participants responded to this question. Percentages based on the entire population (18 participants) are given below in Table 1.

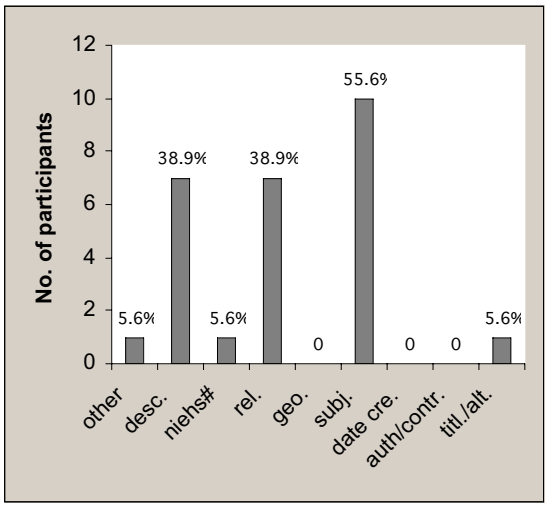
**Table 1. Communication Methods Preferred for Collaborative Metadata Generation**

| Communication Method | Valid Percent selected |
|---|---|
| Face to face | 7 (38.9%) |
| Email | 6 (33.3%) |
| Phone | 2 (11.1%) |
| Web form | 2 (11.1%) |
| Other | 0 (0.0%) |

The preferred communication methods were "face to face" (personal contact) and "electronic mail". No relationship was found between preferred communication methods and desire for (or not for) cataloger assistance.

The last segment of the collaboration survey examined metadata elements that participants would like help generating. The NIEHS metadata schema is an application profile based on the Dublin Core metadata schema, the GEM (Gateway to Educational Materials) (www.geminfo.org) schema, and the NIEHS namespace (Robertson et al. 2001). The NIEHS application profile underlying this study is reported on in this conference's proceedings (see: poster session, Harper et al. 2002). A checklist consisting of descriptive labels for eight NIEHS Dublin Core metadata elements was given to participants, with the option to add "other" elements as desired. (The check-list included "Title/Alternative title", "Author/Contributor", "Date created", "Subject keywords", "Geographic coverage", "Relationship w/other resources", "NIEHS number", and "Writing a description"). Participants were asked to each select three or fewer elements from the checklist. It should be noted that the checklist was purposefully not inclusive but served to prompt participants thinking about their metadata generation experience. To facilitate this process, participants were also encouraged to view the metadata records they produced prior to the collaboration survey, which was reproduced on yellow paper.

Figure 1 graphically compares the selection of each individual metadata element by participants. Results show that more than half of the participants favored cataloger help for "subject keyword" metadata (10 of 18 participants, 55.6% selected this element). Participants also favored cataloger help for "relationship" and "description" metadata to a fair degree, as both of these elements were selected by 7 of the 18




*percentages given past on participant sample of 18 per element.

**Figure 1. Author's Selection of NIEHS Dublin Core Elements for Cataloger Assistance**

participants (an 38.9% selection rate). "Other", "NIEHS number", and "Title/Alternative title" were each selected once. Among the elements not selected by any of the participants for assistance were "Author/Contributor", "Date created", and "Geographic coverage". "Author/Contributor" and "Date created" are first-hand knowledge for authors and not very complicated to generate, so these results make sense. Most NIEHS web pages do not concern a geographic focus, so leaving this element blank may have been an obvious choice.

One-third (6 of 18, 33.3%) of the participants did not select any elements. Four of these participants where four of the six (66.7%) who indicated that they did not think cataloger assistance would be useful during metadata generation. The other two participants who originally indicated that they did not think cataloger assistance would be useful each selected one metadata element where they thought cataloger guidance would be useful: One of these participants selected "other" and identified "new publications" and the other participant selected "description".

A final open-ended question asked participants how they envisioned collaborating with a cataloger. Responses indicated participants' awareness and concern about metadata currency. For example, one participant said that, "as new pages are developed or old pages are modified, program staff would meet with web designer and cataloger to ensure the site is easily accessed by the appropriate audience(s)". Another example is offered by a participant who asked, "how often will this [metadata] be upgraded? (PI [principal investigator] leaves in a few days and new PI arrives in mid-June)". This second example reveals the participants concern for the PI represented in the metadata, not being up-to-date. Several participants commented on preferred communication methods and interest in collaboration, while three participants noted their confusion about the word metadata.

## 7. Discussion of Results

The results of this study provide insight into authors' views on collaborative metadata generation, preferred communication methods for collaborative metadata generation, and types of metadata authors are most likely to seek expert help generating. Moreover, they provide clues about how to expedite the creation of quality metadata through a collaborative model to help build the Semantic Web.

*Authors' views on collaborative metadata generation.* The majority of participants in this study support collaborative metadata generation. They recognized that catalogers have special knowledge and skills—particularly in working with standards. Furthermore, participants recognized that, as resource authors, they too have knowledge valuable to cataloging. Scientists are in the business of generating data. They are avid users of commercial database and often depend on data sets created by other researchers. Data integrity is critical to scientific work; good data furthers research efforts, leads to new discoveries, and advances knowledge. Given these scenarios, it makes sense that the scientists participating in this study demonstrated an understanding of the importance of producing quality metadata in an efficient manner, and supported collaborative metadata generation.

*Preferred communication methods for collaborative metadata generation.* Communication methods are key to any collaborative operation. Participants in this study were equally in favor of both personal (face to face) communication with catalogers and using electronic communication protocols supported by e-mail. The web offers glorious new communication capabilities for disseminating and accessing resources. For example, it's fairly easy to video-conference with colleagues across the globe from the comfort of your own office. Although the results of this study indicated two preferences, it's very likely that these results will change over time, particularly with the introduction of new technologies. Likewise, preferences will change as collaborative partnerships grow, and partners (authors, catalogers, etc.) comfort levels are established. Additionally, different collaborative partnerships and partners will prefer different communication protocols, or work with a combination of methods (for example, "e-mail" and "fact to face" meetings or "telephone"). The researchers in this study advocate that collaborative metadata operations remain open to and test new technologies and various combinations of methods on both a team and

individual level. Related to this is the need to explore human computer interaction (HCI) questions and the design of web forms and tutorials developed to help authors in a collaborative metadata project.

*Types of metadata authors most likely to seek expert help generating.* Participants selected "subject" metadata for cataloger assistance more than any other element, indicating that this element might be the most problematic for authors. Participants' selection of "subject" metadata was further supported by their high selection of the "description" metadata element, which includes *abstracts*, *summaries*, *content notes* and other descriptions with an *intellectual aboutness* (subject) component. Based on these results, subject metadata is an area where experts might focus their attention. Experts could help educate authors through interactive sessions. Experts might also provide tutorials for using the wide-variety of subject tools available on the web, many lack user friendly introductions about the principles of subject analysis. The larger metadata generation experiment referred to above included a metadata tutorial with slides illustrating how to achieve subject *specificity* and *exhaustivity* when creating metadata. The results of the author generated metadata needs to be analyzed and may provide further insight into this issues.

What is perhaps most significant about subject metadata is its relationship to ontology construction and use and the goals of the semantic web. Achieving semantic interoperability and sharing ontological structures are critical for building the Semantic Web (Heflin & Hendler 2000). (Example, see also homepage for ECAI-02 Workshop on Ontologies and Semantic Interoperability: www.informatik.uni-bremen.de/~heiner/ECAI-02-WS/). Underlying this goal is the need to accurately employ metadata schemas and assign ontological terminology, particularly *subject terminology*. Explained another way, without semantic interoperability, there can be no Semantic Web, because agents have no intelligent knowledge structure to roam, make linkages, and complete tasks. Subject metadata is at the core of many ontologies that are being constructed and needs to be studied from a number of different venues.

## 8. Conclusion

This study provides insight into aspects of collaborative metadata generation that may be useful for achieving the Semantic Web. Although the sample size was limited, the results are fairly consistent and provide data for comparing results gathered from additional research in this area. Another contribution of this work is that the research design provides a framework for future studies examining collaboration among authors and experts during metadata creation, as well as for other classes of persons (e.g., professionals and para-professionals).

The goal of the Semantic Web is to support sophisticated tasks, such as planning a vacation. Web agents are essentially problem solvers, in that a person seeks assistance from a web agents, which roams the web to complete a task or provide an answer. One of the major limitations to this simple idea is that there is not nearly enough web content metadata to facilitate sophisticated web agent roaming and task activities. Examining the potential for collaborative metadata generation by drawing upon the expertise of different classes of persons is one way to contribute to remedying this problem—herein is the topic underlying this paper.

Scientific, government, and educational institutions are among leading users of the web technology. Their information is non-proprietary and produced for betterment of society. These agencies have a vested interested in their resources being used for problem solving and in seeing the realization of the Semantic Web. Research needs to further explore options whereby authors and experts in these institutions may effectively collaborate to efficiently generate good quality metadata and contribute to a foundation for the Semantic Web. The results presented in this paper indicate that the authors' surveyed are supportive of a collaborative metadata operation, at least in a governmental institution.

In conclusion, the integration of expert and author generated descriptive metadata can advance and improve the quality of metadata for web content, which in turn could provide useful data for intelligent web agents, ultimately supporting the development of the Semantic Web. The Dublin Core's application in a wide-variety of metadata initiatives and its use by many different classes of persons (experts, authors, web developers, etc.) provides opportunity for collaborative metadata generation involving different classes of persons and in different environments. If such partnerships are well planned and evaluated, they could make a significant contribution to achieving the Semantic Web.

## Acknowledgements

This research was funded in part by OCLC Online Computer Library Center Inc. and Microsoft Research.

The authors would like to acknowledge the following members of the NIEHS metadata team and School of Information and Library Science, University of North Carolina (SILS/UNC) for their help with the collaborative metadata survey research: Ellen M. Leadem (NIEHS Library), Corey Harper (University of Oregon Libraries), Oknam Park (SILS/UNC), Bijan Parsia (UNC/University of Maryland Semantic Web Research Team), Cristina

Pattuelli (SILS/UNC), Nikki Warren (SILS/UNC and NIEHS intern), and Wayne Wiegand (SILS/UNC and NIEHS intern).

## References

Anderson, J.D. & Pérez-Carballo, J. 2001. The nature of indexing : how humans and machines analyze messages and tests for retrieval. Part I. Research, and the nature of human indexing. Information Processing & Management, 37: 231-254.

Dublin Core metadata element set, version 1.1: reference description. 1997. Available at: http://purl.org/DC/documents/rec-dces-19990702.htm.

Duval, E., Hodgins, W., Sutton, & Weibel, S. 2002. Metadata principles and practicalities. D-Lib Magazine, 8(4): http://www.dlib.org/dlib/april02/weibel/04weibel.html.

Dempsey, L., Heery, R., Hamilton, M., Hiom, D., Knight, J., Koch, T., Peereboom, M. & Powell, A. 1997. DESIRE - Development of a European Service for Information on Research and Education (RE 1004 ): http://www.ukoln.ac.uk/metadata/desire/overview/.

Greenberg, J. 2002. Metadata and the World Wide Web. Encyclopedia of Library and Information Science. New York: Marcel Dekker, Inc. 72: 344-261.

Greenberg, J., Pattuelli, M.C., Parsia, B. & W.D. Robertson. 2001. Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. Journal of Digital Information (JoDI), 2(2): http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/.

Harper, C., Greenberg, J., Robertson, W.D. & Leadem, E. (2002). Abstraction versus implementation: issues in formalizing the NIEHS application profile. In DC-2002: Metadata for e-Communities: Supporting Diversity and Convergence, Florence, October, 13-17, 2002.

Heflin, J. & Hendler, J. 2000. Semantic Interoperabilty on the Web. Extreme Markup Languages, 2000. http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf

Lassila, O., Berners-Lee, T. & Hendler, J. 2001. The Semantic Web
A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, (May, 2001): http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2

Milstead, J. & Feldman, S. 1999. Metadata: Cataloging by any other name. Online, (January/February): 25-31. http://www.onlineinc.com/onlinemag/OL1999/milstead1.html.

Patel-Schneider, P.F., Horrocks, I. & van Harmelen, F. 2002. Feature Synopsis of the OWL Ontology Language: http://lists.w3.org/Archives/Public/www-archive/2002May/att-0021/01-_owl.html

Roberston, D., Leadem, E., Dude, J. & Greenberg, J. 2001. Design and Implementation of the National Institute of Environmental Health Sciences Dublin Core Metadata Schema. Proceedings for International Conference on Dublin Core and Metadata Applications 2001, October 22-26, 2001, National Institute of Informatics, Tokyo, Japan, pp. 193-199.

Schwartz, C. 2002. Sorting Out the Web: Approaches to Subject Access. Westport, Connecticut: Ablex Publishing. Part of the Contemporary Studies in Information Management, Policies, and Services series by Hernon, P. (Ed.).

Thomas C. & Griffin, L. 1999. Who will create the Metadata for the Internet? First Monday: Peer Reviewed Journal of the Internet: 131.193.153.231/issues/issue3_12/Thomas/index.html.

Weinheimer, J. 2000. How to Keep the Practice of Librarianship Relevant in the Age of the Internet. Vine (Special issue on Metadata, part 1), 116: 14-27.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 53-60*


# Preliminary Results from the FILTER Image Categorisation and Description Exercise

Jill Evans
Institute for Learning and Research Technology, University of Bristol
jill.evans@bristol.ac.uk

Paul Shabajee
Graduate School of Education and Institute for Learning
and Research Technology, University of Bristol
paul.shabajee@bristol.ac.uk

## Abstract

*Although there are now vast numbers of digital images available via the Web, it is still the case that not enough is known or understood about how humans perceive and recognise image content, and use human language terms as a basis for retrieving and selecting images. There is an increasing belief that the difficulties of image management and description should be led and defined by the needs of users and by their information seeking behaviours. The Focusing Images for Learning and Teaching – an Enriched Resource (FILTER) project is investigating, through an online image description exercise, the ways that users describe different types of images. Through analysis of the exercise results, FILTER hopes to obtain an understanding of the ways in which people describe images and the factors that influence their approaches to image description and thus appropriate metadata. Preliminary analysis of data indicates that there is little consensus on the use of terms for image description or on categorisation of images into 'types'.*

## 1. Introduction: Describing and Retrieving Images

As all forms of communication are increasingly transferred via the grammar of the visual, humans are becoming more sophisticated in their ability to recognise and interpret visual meaning and are using visual information to enhance social, cultural or learning activities [1]. The huge global financial investment in the digitisation of analogue images means that there are now immense numbers of diverse image collections available on the Web, and appetite for consumption of visual information continues to grow. However, it is still the case that not enough is known or understood about how humans perceive and recognise image content, and use human language terms as a basis for retrieving and selecting images in vast, complex, heterogeneous online environments [2]. There is a gap in understanding of how humans verbalise visual perceptions and beliefs within paradigms typically associated with text. Consequently, we cannot be sure that the image resources we are creating within the constraints of orthodox description models are actually being found, accessed and used by our target audiences.

The creators of digital image collections, in most cases, intend their collections to be used as widely and as effectively as possible. Therefore, adding as comprehensive and rich a layer of metadata as possible is essential. Metadata is particularly important for images, as without accurate description there cannot be accurate retrieval [3]. An enormous problem, however, is that images, by their very nature – that they are pictorial representations rather than verbal – are difficult to classify in orthodox textual terms. O'Connor [4] notes that images are complex '… by being literally and figuratively unquotable, everlastingly slipping through in the instance of being identified, seized for scrutiny'. Additionally, '… there is no saying just how many words or just which words are required to describe any individual pictures' [5]. Typically, therefore, visual material requires description of a greater depth than textual resources in order to convey an understanding of its complexity and the implicit multiple layers of meaning and content [6].

Interpreting the meaning or essence of an image and translating that into a surrogate form for retrieval and management is a complicated and diffi-

cult process. Description of images is made complex because of the impossibility of agreement on the meaning, or on the most critical aspects represented. Meaning will be constructed differently, and various elements will have greater relevance to different users depending on, for example: the intended use for the image, the user's area of study or research, the user's educational level of ability, his/her level of cultural, social and historical awareness, and so on. Any or all of these factors will provoke a different user reaction to the image and influence perception of what is depicted and what is expressed; clearly, the vocabularies employed by this wide range of users will all be quite different and will be impossible to anticipate by the providers of the digital image collection.

For example, the following image is a histological stain of tissue from a human gland.

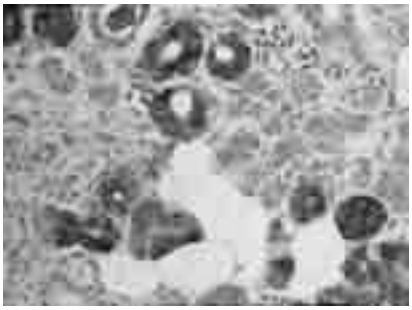


**Image 1. (Copyright the Bristol Biomedical Image Archive)**

The intended audience is doctors, nurses, medical students, and so on. However, the image could equally be of use to an art student looking for inspiration for a textile design. Evidently, the subject content of an image can hold a variety of meanings depending on the primary need of the user but it would be impractical and almost certainly impossible for the digital image collection provider to attempt to describe and classify all these possible contexts and uses.

As Conniss, Ashford and Graham [2], have observed the depth and manner in which an image is catalogued will have consequences for its ability to be retrieved by a potential user. They go on to note that, although comprehensive, in-depth cataloguing can provide multiple and varied entry points for diverse audiences, this level of description is in practice a very time-intensive and ultimately costly process. If images are to be added to a collection regularly, the issue of sustaining such a cataloguing model needs to be addressed.

A book or journal has a generally predictable structured format. Subject indexing of textual materials is usually aided by the availability of several sources of information: preface, title, table of contents, and so on, from which to determine the primary content, aims, scope and purpose. It is normally possible to locate appropriate subject headings to describe the item from within a controlled vocabulary or thesaurus. Cataloguers of visual material typically have no such standard, recognised sources at their disposal, and images can take many varied forms. Howard Besser [7], on the topic of image description writes:

*(of a book) '... Authors and publishers go to great lengths to tell us what this purpose is, citing it in the preface and introduction, on dust covers, and often on the book's cover. Images do not do this. To paraphrase one prominent author speaking of museum objects, unlike a book, an image makes no attempt to tell us what it is about. Even though the person who captured an image or created an object may have had a specific purpose in mind, the image or object is left to stand on its own and is often used for purposes not anticipated by the original creator or capturer'.*

There will usually be little textual information accompanying the image to explain the social, political, historical, religious or cultural context in which it is embedded, or for what purpose the image was created. The image cataloguer may, therefore, have to invest considerable time in research in order to answer these questions before attempting image description.

Most images comprise two aspects of meaning: what they are 'of', that is, what they depict, such as an identifiable person, object or place; what they are 'about', an underlying theme, emotion, message or other abstract concept that is extracted by interpretation. These 'ofness' and 'aboutness' aspects are based on the model developed by Panofsky [8] who described levels of meaning as pre-iconographic (of) and iconographic (about). Krause [9] has used the terms 'hard indexing' and 'soft indexing'. For example, the German Expressionist artist Kathe Kollwitz's charcoal drawing overtly depicts an impoverished mother and her children, but is covertly about the plight of the working classes during the years of the German depression.

**Image 2. (Copyright Brooklyn College History Dept.)**

Far from being a straightforward depiction, this image was intended as a stringent criticism of government policy and beliefs by an artist violently opposed to them. However, without an understanding of the historical and social factors that influenced the creation of this image, the meaning cannot fully be understood. Cataloguing images for this type of conceptual or abstract information need is extremely complex as possibly no two indexers will reach consensus on the subjective qualities or meaning of an image. As Bradfield [10] notes:

*'The problem with the retrieval of visual material is that it evokes concepts related to the reality which it represents. Such concepts are not easily expressed in words but are the 'sought' features of that visual image. Equally, the reality is not always readily expressible in words'.*

It can be seen that determining the focus of an image for indexing poses a huge challenge, both from a depiction and expression perspective. A user reacts to an image on many levels in constructing a feeling for its meaning and connotations. Even images that are abstract and elusive in content are capable of evoking feelings and attempts to communicate those feelings through words. Ornager [11], based on research conducted with image users, suggests that image indexing should indicate what the image depicts (its 'ofness'), what it expresses ('aboutness'), and, additionally, the contexts in which it can be used.

It is generally accepted that the aim of image indexing is not only to provide access based on descriptions of attributes, but also to provide access to useful and logical groupings of images [12]. However, images, unlike verbal representations, share no common hierarchy or taxonomy of definitions and relationships. O'Connor [5] gives as an example a series of images of an elephant, a sheep and a horse, all of which could more broadly be described as animals. A user looking for an image of a horse might also be interested in seeing images of other animals, or of animals that graze, or of mammals, or of four-legged mammals. How can images be indexed and represented to users so as to make these complex but potentially useful relationships more visible and accessible? How can indexers predict what kind of relationships will be valuable to diverse communities of users? [13].

## 2. The FILTER Project

There is an increasing belief that the difficulties of image management and description should be led and defined by the needs of users and by their information seeking behaviours [6], [14], [11], [5]. The Focusing Images for Learning and Teaching – an Enriched Resource (FILTER) project (http://www.filter.ac.uk/) is investigating, through an online image description exercise (http://www.filter.ac.uk/exercise/), the ways that users describe different types of images. About 40 copyright-free images of varying original types (e.g. map, etching, drawing, chart, painting, and so on) and subject content were placed online; individuals from all aspects of tertiary education (but not restricted to these) were invited to participate in the exercise by, firstly, describing the subject content of each image in a series of unlimited keywords, and secondly, describing the type of image (the original type rather than the digital type, as all could legitimately be described as a 'photograph'). Through analysis of the exercise results, FILTER hopes to obtain an understanding of the ways in which people describe images and the factors that influence their approaches to image description. For example: are there particular 'types' of images (e.g. line drawings, graphs, maps) that are easier to describe – and where more consensus is reached? How do users react to images that are more abstract or ambivalent in content compared to images where the content is clear – are more words used to describe ambiguous content? When text is included in an image, does this influence choice of keywords? Is there a difference in the way users from different subject areas approach image description?

FILTER is working with academic image users to develop an exemplar database of image-based learning and teaching materials that demonstrate effective use of images across subject areas and in a range of pedagogical contexts. FILTER has recognised that there are complex issues involved in making these materials and the images embedded within them available in a heterogeneous environment [15]. Both resources and images need to be described and represented in such a way as to encourage users from a specific subject area to look beyond that to examples of image use in other disciplines, which might be relevant. In order to achieve this cross-searching and potential transference of knowledge and expertise, we first need to achieve an understanding of how people perceive and describe images.

## 3. The Image Exercise

It was essential that the images included should be very diverse in 'type', subject content and style of content representation (i.e. degree of clarity/ambiguity). Also important was that the range of images should be typical of those used in different pedagogical contexts. Images were randomly assigned a number from 1-41. On accessing the Web page, participants were presented with a random image accompanied by a questionnaire. Once the questionnaire had been completed and submitted for that image, participants were offered the choice of proceeding to the next image in the sequence. Participants were not permitted to opt out of 'difficult' images or choose which images to describe.

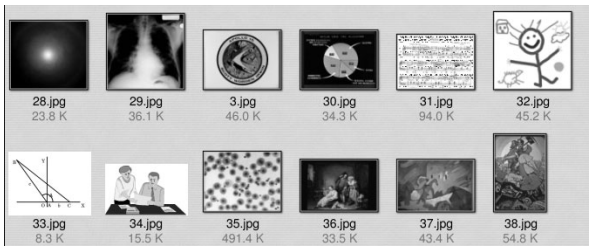



**Image 3. Screenshot of a selection of images used in the exercise (Image copyright NASA, NOAA, FWS, Brooklyn College History Dept., AICT, Bristol BioMed, FILTER)**

### 3.1 The Questionnaire

In addition to adding their descriptions, participants were required to add their area of study, teaching or research and their status (e.g. higher education lecturer, librarian, further education student). The questionnaire was by default anonymous but participants were given the option of adding their contact details should they wish to be involved in future FILTER research (UK Data Protection regulations were a consideration here). From this data we can identify 251 individuals (Table 1). Participants study, work or teach within a diverse range of subject areas, for example: Education, Music, Art and Design, Environmental Science, Medicine, History, Language and Literature, IT, Biology, Psychology, Engineering, Librarianship, Archaeology, Law, Business, Management, and so on.

### 3.2 The Sample

The sample for the exercise was self-selecting but particular groups of potential image users were targeted for publicity. Information about the exercise was sent to a variety of UK and international mailing lists in the education, library, imaging and educational technology fields. The exercise was also published on relevant Web sites, in newsletters and promoted by word of mouth.

**Table 1. Status of participants and numbers participating**

| Status of participant | Number participating |
|---|---|
| Higher Education lecturer | 62 |
| Further Education lecturer | 7 |
| Higher Education student | 16 |
| Further Education student | 5 |
| Researcher | 43 |
| Librarian (HE & FE) | 68 |
| Educational technologist | 18 |
| School teacher | 3 |
| Administrative staff | 8 |
| Technician | 3 |
| Digitisation staff | 4 |
| Other categories | 3 |
| Status not given | 11 |

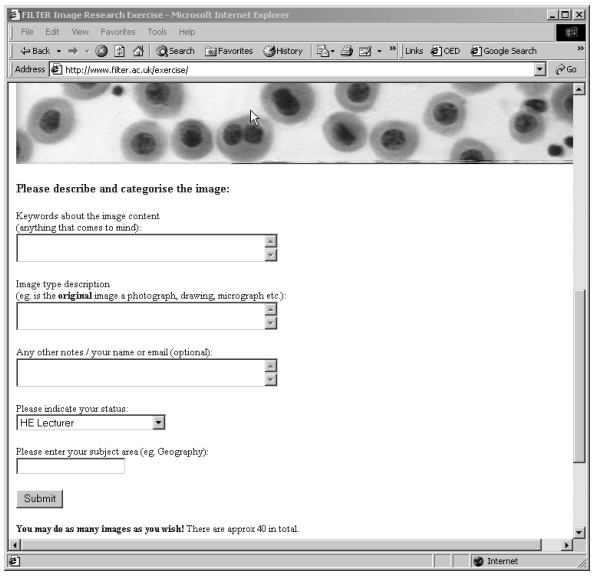



**Image 4. Screenshot of questionnaire**

## 4. Preliminary Results

The aim of the survey is to gain an initial insight into the ways a variety of image users perceive, describe and categorise images. The survey is not intended to be a definitive study in the area, but rather to highlight issues for further, more rigorous, investigation and research. The self-selecting nature of the sample and the self-reporting format of the Web questionnaire do raise questions of validity and unreliability: it is impossible to ascertain to what degree respondents are representative of the community of image users, or whether a full range of expertise in image use – from novice to expert – is present. However, as a basic snapshot of the current level of expertise in the field of visual information description and classification across multiple disciplines, we believe the survey has great potential relevance. The substantial proportion of respondents who provided their contact details indicates that preliminary findings can be followed up via interview or other in-depth questioning.

### 4.1 Analysis

An initial analysis of the data has been conducted by taking the submissions from the Web form, entering the data into a relational database and querying the database directly or producing output for more detailed analysis in the statistical analysis package SPSS and qualitative analysis software ATLAS-ti

where appropriate. At the time of this preliminary analysis there had been 1150 responses.

*The 'Other' category contains a range of roles which, in cases where respondents provided additional details, we have been able to further categorise (see Table 1).*

Figure 2 shows that the number of responses for each image was fairly evenly distributed as expected due to the random presentation of the images to the participants.

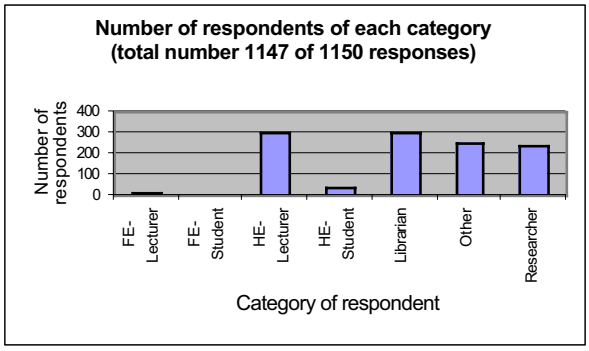



**Figure 1. Number of respondents of each category**

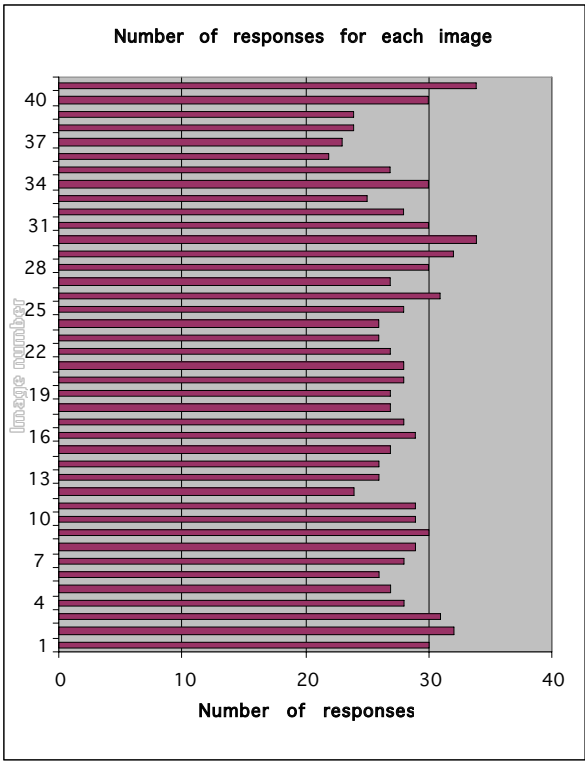
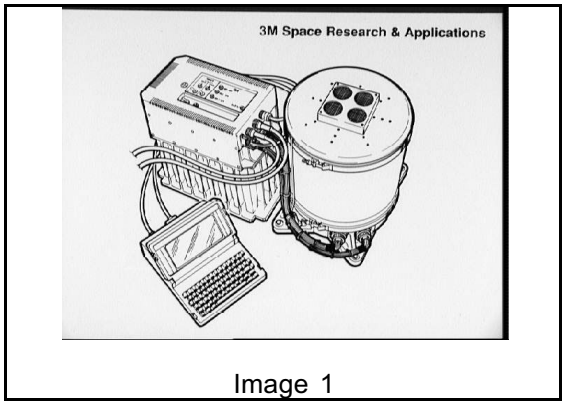



**Figure 2. Number of responses for each image**

**Table 2. Terms used to describe the type of image for Image 1 (Image copyright NASA)**




| Term or Phrase | Number of times used |
|---|---|
| drawing | 16 |
| line drawing | 2 |
| technical drawing | 1 |
| artwork | 1 |
| technical sketch of overall arrangement of kit | 1 |
| sketch | 1 |
| pen-and-ink drawing on paper | 1 |
| line illustration for technical manual | 1 |
| engineering drawing | 1 |
| drawing probable computer generated | 1 |
| computer-generated image | 1 |
| computer generated image | 1 |
| two tone drawing | 1 |

**4.2 Categorisation of Image Types**

A total of 391 terms and phrases were used to categorise the types of image in the exercise. This number includes terms and phrases with obvious spelling errors and hyphenated alternatives.

Table 3 shows the terms and phrases used more than once to describe the type of image and the number of times it was used across the whole collection. This illustrates a number of interesting points: 1) it seems that synonyms occur frequently, e.g. photo/photograph and computer image/computer-generated image/digital image 2) qualifiers are used extensively, e.g. colour, black and white, computer generated, annotated 3) misspellings occur, e.g. 'satelite'. This is more common than shown here as many spelling mistakes occurred only once 4) some contextual terms are used, e.g. scientific, and 5) participants use terms such as 'don't know' and 'probable' as indicators that they are unsure how to categorise the image. We plan to use this data in combination with earlier work to develop a set of vocabularies and relationships that model how participants of this and other studies have described image types.

**Table 3. Terms and phrases used more than once to describe the type of image and the number of times it was used (for the whole set of images)**

| Term or Phrase | N | Term or Phrase | N |
|---|---|---|---|
| photograph | 23 | annotated map | 2 |
| drawing | 21 | colour drawing | 2 |
| photo | 17 | computer generated drawing | 2 |
| painting | 10 | computer generated image | 2 |
| colour photograph | 8 | computer image | 2 |
| print | 7 | don't know | 2 |
| diagram | 6 | engraving | 2 |
| black and white | 5 | figure | 2 |
| chart | 5 | logo | 2 |
| micrograph | 5 | map drawing | 2 |
| sketch | 5 | microscope image | 2 |
| black and white photograph | 4 | microscopic image | 2 |
| computer-generated image | 4 | photograph (colour) | 2 |
| graph | 4 | photograph of graphic | 2 |
| illustration | 4 | photographs | 2 |
| line drawing | 4 | picture | 2 |
| artwork | 3 | printed map | 2 |
| b&w photograph | 3 | printout | 2 |
| cartoon | 3 | satelite image | 2 |
| computer generated | 3 | scan | 2 |
| lithograph | 3 | scanned image | 2 |
| map | 3 | scientific graph/ diagram | 2 |
| microscope slide | 3 | technical drawing | 2 |
| photo of chart | 3 | trace | 2 |
| aerial or satellite photograph | 2 | transparency | 2 |
| aerial photograph | 2 | | |

For each image a clear pattern emerged, as illustrated in Table 2 (for the case of image 1). A small number (generally 1-4) of terms or phrases were used by a large proportion of participants, and a larger number of terms were used only once, i.e. by only one participant in each case. This implies that for each image there is a small set of 'image types' that are most commonly used.

Figure 3 shows the number of distinct terms or phrases used to describe each image type in the collection. Initial analysis indicates that there are underlying patterns in this data. For example, images 6, 11, 18 and 26 all have 'number of types of image' of 5 or less and in each case the most popular term to describe that image type is 'photograph', whilst this was not the case for any of those images with more than 15 terms or phrases.

A more detailed analysis will, we hope, provide insight into exactly what characteristics of an image participants have focused on in order to arrive at a categorisation type.

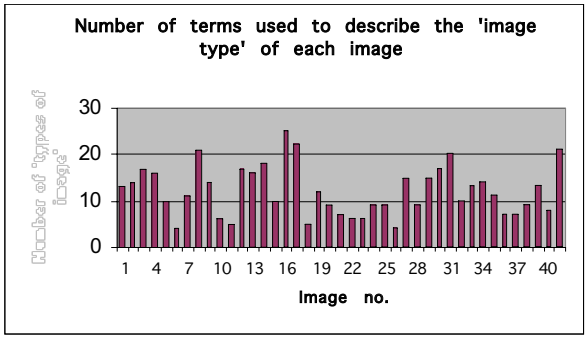



**Figure 3. Number of terms used to describe the image type of each image**

### 4.3 Describing the Images

The description of images is more complex to analyse. Initial analysis described here uses only individual words although participants frequently used prose to describe the images. Thus the meanings that can be derived from that prose are lost in this analysis. Figure 4 shows the number of terms used per image. For all images it appears that the distribution of the number of terms is very flat. Figure 5 shows the number of times a word was used to describe image number 1.

For image 1 there are 130 terms in total used by 30 participants; the most commonly used is 'computer', used 18 times, followed by 'equipment', used 11 times. There are 54 terms that have been used only once. This pattern is similar to that for the image type data above but with many more terms and a longer tail. It is more problematic to analyse these with respect to categorising the words and thus identifying common themes. This more detailed analysis will be conducted over the next months, however, we describe our initial analysis and findings below.

We are currently analysing the descriptive terms used for each image in detail. Table 2 shows the words used to describe image 1. Figure 6 shows patterns of the co-occurrence of words in descriptions for image 1. The terms joined by lines were terms that co-occurred more than once in the descriptions and the thick lines indicate those that co-occurred more than three times. Clearly, for image 1 the dominant concepts are: computer, equipment, control/controlled, space and research.

Preliminary findings indicates that, on further analysis, it should be possible to categorise the words and phrases in terms of the nature of the description, for instance: shape of content, colour, object, person, historical/temporal. We should also be able to judge the extent to which the inclusion of text in an image influences amount and choice of words. Such analysis will, we believe, provide a means of categorising the characteristics in the images that participants have used to describe the images and of gaining insight into the elements or facets of the images that they use to make their choices of terms.

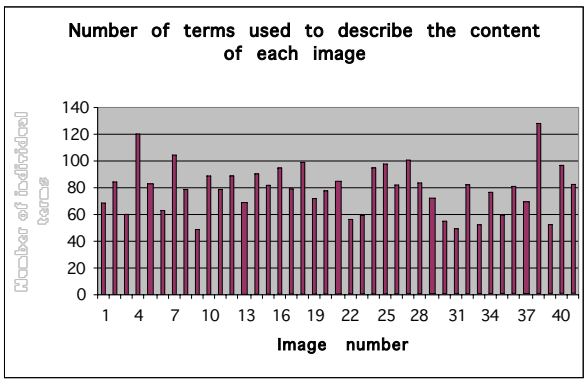



**Figure 4. Number of terms used to describe the content of each image**

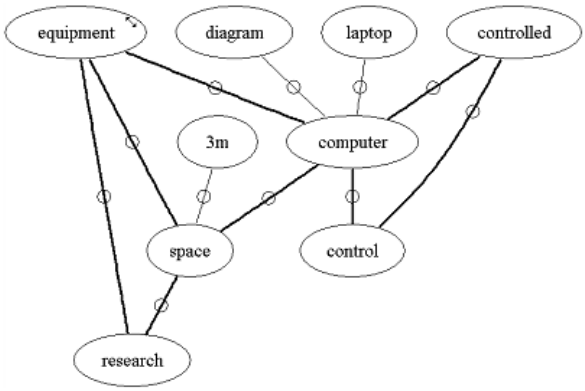



**Figure 6. Graph of co-occurrences for image 1 produced with Graphvis software [16]**

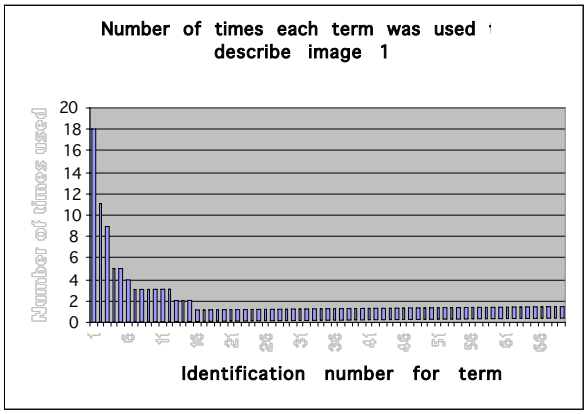



**Figure 5. Number of times each term was used to describe image 1**

## 5. Discussion and Moving Forward

The findings, even for this early analysis, seem to point to the advantages of, and need for, the use of controlled vocabularies and thesauri to overcome common spelling variations and errors. Current image metadata standards do not seem to provide vocabularies for representing the full range of 'image type' descriptors used by the participants. Of course, it is not clear that such a categorisation is necessary, or would be beneficial. However, if a digital image collection contains tens of thousands of images of maps, and a user is specifically seeking paintings of maps, then certainly some form of metadata standards would be required to enable such a search to be carried out efficiently.

On the basis of the data presented in this paper, one possible structure for an image type categorisation system would be to define core types (e.g. photograph, painting, map), with an optional set of qualifiers for providing additional meaning based on visual aspects of the image. These might be, for example, colour/black & white, mode of generation/technique/process, angle/view, scale, spatial/temporal. Clearly there is some potential here for the development of a refined set of qualifiers for the DCMI Type Vocabulary [17].

The Getty Art and Architecture Thesaurus (http://www.getty.edu/research/tools/vocabulary/aat/) is a highly authoritative vocabulary in the broad arts fields; FILTER suggests that it would be useful to develop a broad categorisation of image types that could be used across a diverse range of subject areas and that is specifically focused on the needs of the tertiary education community. The results of the exercise clearly demonstrate that there is a lack of consensus on the process of recognising and categorising images by type. It is possible that this lack of agreement acts as a barrier in the successful retrieval and use of images and that this issue could be addressed by the establishment of a common vocabulary.

There is clearly very large diversity in the terms used to describe the content of the images in this sample. While dominant terms exist for each image, they represent only a small percentage of the terms used and are not used by all participants. While thesauri can clearly facilitate the mapping of terms to any descriptive metadata, it is not clear from our current analysis that such a mapping would meet the needs of the participants. Once performed, these analyses should help to shed light on this issue.

Other analyses that we are conducting are based on the identification of patterns in the data related to, for instance, the role or "subject area" of the participant and their classification of both image type and content. For example, we might hypothesise that participants from different subjects will describe the content in different ways. It is hoped that this work will enable FILTER and other projects develop effective means of helping those developing online image collections for educational purposes to provide their users with tools to retrieve images.

## References

[1] Kress, G. and van Leeuwen, T., *Reading images: the grammar of visual design*. Routledge. London, 1996.

[2] Conniss, L.R., Ashford, A.J. and Graham, M.E., *Information seeking behaviour in image retrieval: VISOR I final report*. (Library and Information Commission Research Report: 95), Newcastle: Institute for Image Data Research, 2000.

[3] Rorvig, M.E., Turner, C.H. and Moncada, J., "The NASA image collection visual thesaurus", *Journal of the American Society for Information Science*, 50 (9), 1999, 794-798.

[4] O' Connor, B.C., "Access to moving image documents: background concepts and proposals for surrogates for films", *Journal of Documentation*, 41 (4), 1985, 209-220.

[5] O' Connor, B.C., "Pictures, aboutness, and user-generated descriptors", *SIGVIS Newsletter* [online], 1 (2). 1996, Available from: http://www.asis.org/SIG/SIGVIS/b_oconnor.html [Accessed 28 June 2002].

[6] Turner, J.M., "A typology for visual collection", *Bulletin of the American Society for Information Science* [online], 25 (6). 1999. Available from: http://www.asis.org/Bulletin/Aug-99/turner.html [Accessed 28 June 2002].

[7] Besser, H., Visual access to visual images: the UC Berkeley Image Database Project. *Library Trends*, 38 (4), Spring, 1990, 787-798.

[8] Panofsky, E., "*Meaning in the visual arts: papers in and on art history*", Garden City, N.Y.: Doubleday Anchor, 1955.

[9] Krause, M. G., "Intellectual problems of indexing picture collections", *Audiovisual Librarian*, 14 (2), 1988, 73-81.

[10] Bradfield, V., "Slides and their users: thoughts following a survey of some slide collections in Britain", *Art Libraries Journal*, Autumn, 1977, 4-21.

[11] Ornager, S., "Image retrieval: theoretical analysis and empirical user studies on accessing information in images", *Proceedings of the 60th ASIS Annual Meeting*, 34, 1997, 202-211.

[12] Layne, S. S., Some issues in the indexing of images. *Journal of the American Society for Information Science*, 45 (8), 1994, 583-588.

[13] Shabajee, P., "Primary multimedia objects and educational metadata: a fundamental dilemma for developers of multimedia archives", *D-Lib Magazine* [online], 8 (6). 2002, Available from: http://www.dlib.org/dlib/june02/shabajee/06shabajee.html, [Accessed 28 June 2002].

[14] Jorgensen, C., "Attributes of images in describing tasks", *Information Processing & Management*, 34 (2/3), 1997, 161-174.

[15] Evans, J., Conole, G. and Youngs, K., "FILTER: Focusing Images for Learning and Teaching - an Enriched Resource". In: Kennedy, G. et al., eds., *Meeting at the crossroads: proceedings of the 18th annual conference of the Australian Society for Computers in Learning and Tertiary Education (ASCILITE), Melbourne, 9-12 December 2001*. Melbourne: The University of Melbourne, 2001, 189-196.

[16] AT&T, *Graphviz – open source graph drawing software* [online]. 2002, Available from: http://www.research.att.com/sw/tools/graphviz/ [Accessed 28 June 2002].

[17] DCMI. "*DCMI Type Vocabulary*". Available from: http://dublincore.org/documents/2000/07/11/dcmi-type-vocabulary/ [Accessed 28 June 2002].

Paper Session 2

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 63-69*


# Building Educational Metadata Application Profiles

Norm Friesen
Athabasca University
norm@netera.ca

Jon Mason
*education.au limited*
jmason@educationau.edu.au

Nigel Ward
*education.au limited*
nigel@thelearningfederation.edu.au

## Abstract

*Metadata schemas relevant to online education and training have recently achieved the milestone of formal standardization. Efforts are currently underway to bring these abstract models and theoretical constructs to concrete realization in the context of communities of practice. One of the primary challenges faced by these efforts has been to balance or reconcile local requirements with those presented by domain-specific and cross-domain interoperability. This paper describes these and other issues associated with the development and implementation of metadata application profiles. In particular, it provides an overview of metadata implementations for managing and distributing Learning Objects and the practical issues that have emerged so far in this domain. The discussion is informed by examples from two national education and training communities – Australia and Canada.*
**Keywords:** *application profile, educational metadata, interoperability, CanCore, LOM, Learning Federation.*

## 1. Introduction

With the recent approval of the Learning Object Metadata (LOM) data model as a standard by the IEEE (IEEE 2002) and of Dublin Core's status as a NISO standard (DCMI 2001), metadata models have achieved a stability and level of community commitment requisite to their implementation in the form of application profiles and supporting infrastructure. The consensus represented and codified in these standards provides implementers and developers with a solid foundation for creating metadata infrastructures to meet the needs of national, regional and local educators and learners. Given the necessarily abstract nature of these standards, the task of adapting them to meet the specific and concrete needs of these stakeholders requires interpretation, elaboration, extension, and in some cases, the simplification of their syntax and semantics.

The DCMI and LOM communities started addressing these issues via a shared workplan outlined in the Ottawa Communiqué (Ottawa 2001). In accordance with this plan, these communities subsequently released the important "Metadata Principles and Practicalities" paper (Duval et al. 2002), which emphasized that this work of adaptation is best undertaken through the definition of application profiles:

> An application profile is an assemblage of metadata elements selected from one or more metadata schemas and combined in a compound schema. Application profiles provide the means to express principles of modularity and extensibility. The purpose of an application profile is to adapt or combine existing schemas into a package that is tailored to the functional requirements of a particular application, while retaining interoperability with the original base schemas. (Duval et al. 2002).

However, it is our common experience that the challenge of retaining interoperability with "original base schemas" – and with other related application profiles – is a non-trivial matter. Both adaptation and interpretation play important roles in the process of profiling metadata for the needs of particular projects and communities. As this paper will illustrate, these needs and requirements are also shaped in complex and subtle but significant ways by the policy

and cultural environments in which these projects and communities exist.

The discussion that follows is largely focused on the development of two application profiles:

1) The Le@rning Federation Application Profile (TLF 2002), which combines elements from LOM, Open Digital Rights Language, and accessibility statements; and
2) CanCore (CanCore 2002), a subset of LOM elements, focusing on best practices for element and vocabulary semantics.

Both the Australian and Canadian profiles have been developed in response to unique sets of goals and requirements. Although both profiles emerge primarily from activities in the public education sector, the substantial differences separating these sectors in Australia and Canada – as well as a number of other factors – has resulted in significant differences in emphasis.

Despite these differences, these two profiles demonstrate surprising commonality in terms of their guiding principles and underlying assumptions. The authors hope that this commonality might also inform future work for DC-Education.

The metadata infrastructure under development in both the Australian and Canadian communities is presented in summary form. This is followed by an overview of important practical issues addressed by the projects during creation of their metadata application profiles. An example of how each profile has integrated the work of both the Dublin Core and IEEE LOM communities will be provided. Finally, the fundamental, underlying similarities between the profiles are highlighted.

## 2. The Le@rning Federation Application Context

The Le@rning Federation (TLF) is a five year initiative aimed at developing a shared national pool of quality online learning content for Australian schools within a framework that facilitates distributed access. It has been co-funded within a policy context developed in collaboration between the Australian Commonwealth government and State and Territory education authorities and focused on the strategic importance of fostering online culture in school education.

In context of this collaborative framework, metadata plays a pivotal role. It is required to support the access, search, selection, use, trade and management of Learning Objects. The Le@rning Federation has addressed its metadata requirements through the development of an application profile that combines or references a number of metadata schemes or namespaces.

Development of TLF metadata has been primarily guided by principles of interoperability and pragmatism. The project recognised that adoption of international metadata standards was critical for achieving interoperability between software used to create, manage, distribute, and use learning objects. It also recognised that adoption of metadata standards should not compromise the ability of school education systems and sectors to achieve their own educational priorities. Working within the tensions between adoption of international and national standards and the pragmatic solutions required for The Le@rning Federation has been a challenging and exciting aspect of the project.

A key shared perspective of the Australian k-12 authorities has been the recognition that optimisation of the *learning value* of digital Learning Objects is fundamental in establishing interoperable metadata specifications for TLF. In other words, it is important for the technology to accommodate learning outcomes and curriculum frameworks, rather than requiring these frameworks to be adapted to technical requirements and limitations. Online content in TLF is being designed and developed in the form of Learning Objects that can be deployed in multiple settings. TLF defines Learning Objects as components of online content (animations, video clips, texts, URLs or sequences of such assets) that have educational *integrity*. That is, they possess educational value independent of any one application or context. Learning Objects with educational integrity can be identified, tracked, referenced, aggregated, disaggregated, used and reused for a variety of learning purposes. Such Learning Objects are developed to function both as discrete entities and as aggregate objects contributing to the achievement of particular learning outcomes.

Schools will access TLF online educational content within a framework of distributed access to State and Territory gateways. TLF will provide access to online educational content via a repository called the 'Exchange'. Education systems will retrieve online educational content from the Exchange and distribute Learning Objects through their online systems. The education systems will also provide Learning Object manipulation tools and e-learning environments required by schools.

It is also important to highlight a broader context. With regard to the application of metadata for educational purposes EdNA (Education Network Australia) developed its first (DC-based) schema for the purposes of resource discovery in 1998. At the time, this represented a hard-won consensus among state and territorial education authorities. However, as both internal requirements and international e-learning specifications developed and changed the importance of referencing work done by others (such as the IMS Global Learning Consortium) while also leveraging work already done in EdNA became increasingly clear. It also became clear that managing learning objects would require metadata for functions other than discovery (i.e. requiring reference to the LOM along with other metadata specifications).

## 3. Le@rning Federation Profile Overview

TLF metadata application profile has been developed in recognition of the fact that existing metadata element sets met some of TLF requirements, but no single element set would be capable of meeting them all. Consequently, The Le@rning Federation Metadata Application Profile (TLF 2002) has taken elements from different metadata specifications or namespaces:

- Dublin Core Metadata Element Set, v1.1 (DCMES 1999);
- Dublin Core Qualifiers, (2000-07-11) (DCQ 2000);
- EdNA Metadata Standard, v1.1 (EdNA 2000); and
- IEEE Learning Object Metadata Standard, draft v6.4 (IEEE LOM 2002).

Some TLF requirements were not met by any standard. For this reason, TLF has also defined new metadata elements. All of the elements comprising TLF metadata are grouped into five categories:

The **management** category groups the information related to both the management and discovery of the digital asset as a whole. It contains some common descriptive elements as well as lifecycle and contribution information

The **technical** category groups the technical requirements and characteristics of the digital asset. For example, it contains information on the file types, software and hardware requirements of the digital asset.

The **educational** category supports description of the educational integrity of a Learning Object and includes elements for describing:

- the object's curriculum topic;
- the potential learning outcomes supported by the object;
- teaching methods for presenting the material; and,
- the intended audience for the object.

The **rights** category groups the intellectual property rights and conditions of use of the digital assets. To place a pool of legally reusable educational material within the reach of all Australian students and teachers requires it to be managed in a way that negotiates and provides agreed reimbursement to owners of intellectual property and that facilitates the creation, trade and usage of online content. To achieve this, TLF curriculum content will need to meet relevant statutory and contractual obligations. TLF metadata contains support for digital rights management by including both text and Open Digital Rights Language (ODRL) statements (ODRL 2002).

The **accessibility** category incorporates an Accessibility Specification developed by the TLF that conforms to Commonwealth laws concerning accessibility. The Specification aims to ensure that online resources and services are inclusive of a range of teaching and learning capacities, contexts and environments. It affirms policy commitments by State and Territory education systems to *inclusive* educational provision. TLF metadata contains support for describing the accessibility of Learning Objects in terms of W3C Web Accessibility Checkpoints and TLF-defined learner accessibility profiles.

## 4. CanCore Profile Context

In contrast to many application-profiling activities, the CanCore Learning Object Metadata Application Profile (or simply CanCore) was *not* developed in response to any single project or undertaking. Instead, this profiling initiative was established in November 2000 to address asset management and resource discovery issues common to a number of e-learning projects sponsored by both federal and provincial governments. These include:

- the BELLE (Broadband-Enabled Lifelong Learning Environment) project, a $3.4 million shared-cost initiave funded under the federal government's Industry Canada department. BELLE's objective has been to develop a prototype educational object repository.
- the POOL (Portal for Online Objects for Learning) project, a Pan-Canadian effort also funded primarily by Industry Canada. This initiative has been developing a distributed learning content management infrastructure based on a peer-to-peer architecture.
- CAREO (Campus Alberta Repository of Educational Objects), a project supported by provincial (Albertan) sources and by Industry Canada that has its primary goal the creation of a searchable, Web-based collection of multidisciplinary teaching materials for educators across Alberta.
- The LearnAlberta Portal, a project undertaken by the department of education in the province of Alberta to provide modular, reusable learning resources integrated with provincial k-12 curricula and objectives.

It is worth noting that these projects span both the higher education and k-12 educational domains, with some focusing on the needs of a single province, and others addressing the requirements of users across all the Canadian provinces and territories. A similar heterogeneity is reflected in the institutions which have hosted and otherwise supported CanCore profiling activity. These include TeleEducation, an arm of the New Brunswick provincial government, the Electronic Text Centre at the University of New Brunswick, as well as the University of Alberta, and Athabasca University.

The support of CanCore by such a broad variety of institutions and projects reflects the shared commitment of these organizations to common set of needs and requirements. Many of these shared requirements are shaped by the highly decentralized nature

of Canadian educational policy. Education in Canada falls under exclusively provincial and territorial jurisdiction. Besides forbidding any federal involvement in education administration or delivery, Canadian policy also encourages education to reflect and sustain a multiplicity of languages and cultures — extending well beyond English and French to include aboriginal, Slavic, Asian and other languages and cultures. (Such policies are, in part, responsible for the diversity of projects listed above, and are perhaps also reflected in these projects' emphasis on infrastructure rather than content).

While explicitly requiring a diversity of educational contents and administrative structures, such an environment also has the effect of defining a common set of values and concerns for those developing educational technologies in Canada. Within this context, means of ensuring cultural and linguistic neutrality and adaptability are understood as mandated requirements rather than being perceived simply as desirable virtues. At the same time, these values and concerns are informed by an acute awareness of Canada's relatively small size as a market for content and Internet technologies, as well as its proximity to the world's largest purveyor of these and other commodities. Together, these factors provide a strong inducement for collaboration and cooperation to protect interests of diversity and adaptability. It is therefore not surprising that CanCore was initiated by the projects mentioned above "to ensure that educational metadata and resources can be shared easily among its users as effectively as possible with others across the country" (Friesen et al. 2002).

## 5. CanCore Profile Overview

Given the diversity of projects and players behind the creation of CanCore, it seems natural that this metadata initiative would focus on bringing these stakeholders together under the banner of a single consensual artefact. This artefact is represented by what is now the IEEE LOM standard; and the CanCore initiative began by identifying a subset of LOM elements that would be of greatest utility for interchange and interoperation in the context of a distributed, national repository infrastructure. The CanCore element set is explicitly based on the elements and the hierarchical structure of the IEEE LOM, but CanCore has sought to significantly reduce the complexity and ambiguity of this specification. In keeping with this approach, CanCore has developed extensive normative interpretations and explications of the metadata elements and vocabulary terms included in its "consensual subset" of LOM elements. This work of interpretation and simplification is featured in the CanCore Learning Object Metadata Guidelines (Fisher et al. 2002), a 175-page document distributed at no cost from the CanCore Website.

In this work, CanCore can be seen to take its cue from a definition of application profiles that precedes ones more recently referenced. Instead of "mixing and matching" elements from multiple schemas and namespaces (Heery, Patel 2002), it presents "customisation" of a single "standard" to address the specific needs of "particular communities of implementers with common applications requirements" (Lynch 1997).

The CanCore application profile comprises some 36 "active" or "leaf" IEEE LOM elements. These elements were chosen on the basis of their likely utility for interchange and interoperation in the context of a distributed, national repository infrastructure. Compared to the elements comprising TLF, the CanCore elements are focused fairly exclusively on resource discovery. Those dealing with rights management and educational applications are kept to an effective to a minimum. This emphasis on resource discovery might also be understood as a result of the heterogeneity of the community CanCore is serving. For example, to accommodate the diverse curriculum and learning outcomes schemes and hierarchies developed separately for k-12 education by each Canadian province, CanCore references and explicates the use of almost all of the LOM Classification element group. By way of contrast,, the TLF profile is able to go far beyond identifying generic placeholder elements, and specifies both specialized elements and vocabularies for learning "strands", "activities", "design" and "content/concepts". Moreover, approaches to both educational application and rights management often vary considerably even *within* the projects and jurisdictions served by CanCore. Consequently, in further contradistinction to TLF, CanCore has not sought out a role in achieving consensus or coordination between *between* Canadian projects on these matters.

## 6. Application Profile Implementation Issues

### 6.1. The Le@rning Federation: Unifying Metadata Information Models

The Le@rning Federation Metadata Specification draws elements primarily from both IEEE LOM and Dublin Core. However, these metadata schemas use different information models for defining and constraining the function of their metadata elements. Unifying these information models has thus been a challenging part of developing the application profile.

The Dublin Core elements are described with an information model based on the ISO/IEC 11179 standard for the description of data elements (ISO 11179). Each element is described using the following ten elements:

- Name – The label assigned to the data element.
- Identifier – The unique identifier assigned to the data element.

- Version – The version of the data element.
- Registration Authority - The entity authorised to register the data element.
- Language – The language in which the data element is specified.
- Definition – A statement that represents the concept and essential nature of the data element.
- Obligation – Indicates if the data element is required to always or sometimes be present.
- Datatype – Indicates the type of data that can be represented in the value of the data element.
- Maximum Occurrence – Indicates any limit to the repeatability of the data element.
- Comment – A remark concerning the application of the data element.

IEEE LOM uses a different set of attributes for describing its elements. Each IEEE LOM element is described using the following attributes:

- Name – The name by which the data element is referenced.
- Explanation – the definition of the data element.
- Size – The number of values allowed.
- Order – Whether the order of values is significant.
- Example – An illustrative example.
- Value space – The set of allowed values for the data element – typically in the form of a vocabulary or a reference to another value space.
- Data type – Indicates whether the values come from an IEEE LOM defined datatype.

The Le@rning Federation application profile has adopted the attributes used by Dublin Core for describing its metadata elements. To incorporate the IEEE LOM element definitions into TLF metadata, the element definitions were recast using the ISO 11179 attributes. In most cases, the mapping was obvious: IEEE Name to ISO Name, IEEE Explanation and IEEE Example to ISO Definition, IEEE Size to ISO Maximum Occurrence, IEEE Data type to ISO Datatype.

The IEEE Order attribute was abandoned because ordered elements were not a requirement for TLF application.

Information in the IEEE Value space attribute was incorporated into the ISO Datatype attribute in TLF definition. It is interesting to note that the IEEE Value space attribute corresponds closely to the Qualified Dublin Core notion of value encoding schemes. In Qualified Dublin Core, encoding schemes identify structure that aids interpretation of an element value. These schemes include controlled vocabularies and formal notations or parsing rules.

Dublin Core elements live in a "flat" space where each element directly describes the one identified resource. IEEE LOM elements, however, live in a "hierarchical" space. Some elements are aggregates of sub-elements. Aggregates do not have values directly; only data elements with no sub-elements have values directly. The sub-elements describe attributes of the aggregated element, rather than the resource directly. For example, the IEEE LOM Relation.Resource aggregation has two sub-elements: Relation.Resource.Identifier and Relation.Resource.Description. These two sub-elements describe the Relation.Resource aggregate rather than the resource being described by the metadata record as a whole.

The hierarchical structure of the IEEE LOM presents a wide range of expressive possibilities. However, such a structure is difficult to integrate with the Qualified Dublin Core notion of element refinements. Element Refinements make the meaning of an element narrower or more specific. A refined element shares the meaning of the unqualified element, but with a more restricted scope. A client that does not understand a specific element refinement term should be able to ignore the qualifier and treat the metadata value as if it were an unqualified (broader) element.

Within the Le@rning Federation application context, it was decided that the IEEE LOM Coverage element should be refined using the Dublin Core Spatial and Temporal element refinements. These element refinements were incorporated into the IEEE LOM aggregation model as sub-elements of the coverage element. This allows distinction between spatial and temporal coverage, but does not meet the Dublin Core requirement that a refinement can be treated as if it were the broader element.

### 6.2. CanCore: Data Model Explication and Simplification

An illustration of CanCore's emphasis on element and vocabulary semantics is provided by its interpretation of the IEE LOM element titled "Learning Resource Type". The discussion of this element provided in the CanCore Metadata Guidelines is also illustrative of CanCore's reference to Dublin Core semantics and best practices as normative guides. In addition, the issues presented by this element and its associated vocabulary also provide evidence of the challenges of facilitating resource for specifically educational resources —and of the need for semantic refinement for even the most rudimentary implementation and interoperability requirements.

The IEEE LOM standard describes the Learning Resource Type element simply as "Specific kind of learning object. The most dominant kind shall be first". The vocabulary values recommended for this element are: "Exercise, simulation, questionnaire, diagram, figure, graph, index, slide, table, narrative text, exam, experiment, problem statement, self assessment, and lecture. In order to provide further guidance on the meaning of these sometimes ambiguous terms, the document refers implementers to the usage histories of the Oxford English Dictionary and to existing practice: "The vocabulary terms are defined as in the OED:1989 and as used by

educational communities of practice". As a final clarification, the data model document also provides a mapping of this LOM element to the "DC.Type" element from the unqualified Dublin Core element set.

In its metadata guidelines document, CanCore interprets these relatively sparse and ambiguous normative indications as follows:

> [The normative information provided by the IEEE LOM] leads to 2 possible approaches, the second of which is recommended by CanCore:
>
> 1. Use the DC Type vocabulary as is (Collection, Dataset, Event, Image, Interactive Resource, Service, Software, Sound, Text), or extend it for the various media in a collection. In each case, the vocabulary is seen as designating a media type, format or genre, relatively independent of the educational purpose or application to which it is put to use. An example of an extended form of the DC recommended vocabulary is provided at http://sunsite.berkley.edu/Metadata/structuralist.html. The fact that this element is [indicated to be] equivalent to DC Type would justify this approach. However, this approach raises the question: How is this element [indicative of a learning resource type]?
> 2. Use or develop a vocabulary that addresses learning very specifically and directly, and the ways that resources can be applied for particular educational purposes. This should occur relatively independently of the actual media type that the resource represents. For example, a text document or interactive resource could be a quiz, an exercise or activity, depending on the way it is used, and the way these educational applications are defined. An example of this type of vocabulary is provided by EdNA's curriculum vocabulary: "activity", "assessment", "course curriculum/syllabus", "exemplar", "lesson plan", "online project", "training package", "unit/module".
>
> The vocabulary values [recommended in the IEEE LOM] seem to conflate these two approaches, including values that indicate media type or format (Diagram, Figure) and values indicating educational application (exam, questionnaire, self-assessment). (Fisher et al. 2002).

In the CanCore guidelines document, this discussion is followed by references to recommended vocabularies developed to designate learning resource types in the context of other projects, as well as multiple text and XML-encoded examples and technical implementation notes. Similar documentation is provided for all of the IEEE LOM elements included in the CanCore subset.

By thus combining best practices from existing data models, implementations and application profiles, and by explicating its own normative decisions, CanCore hopes to provide significant direction and assistance to those making decisions about educational metadata – whether they be administrators, implementers, metadata record creators, or developers of other application profiles. In doing so, Cancore leverages semantic consensus already developed in the Dublin Core community (and elsewhere) to promote semantic interoperability among projects referencing the IEEE LOM, and also to work toward cross-domain interoperability through mutual reference to the DC data model.

## 7. Application Profile Commonalities

In discussing in broad terms contexts and experiences associated with the development of our respective application profiles, some commonly identified principles have emerged. It is clear that both profiles have been developed differently, in response to the requirements of contexts. However, it is hoped that the experience of their development will inform ongoing efforts within educational communities that are developing and implementing metadata schema for resource description and management purposes.

### 7.1. Respecting Existing Practice for Semantic Interoperability

The development of both TLF and CanCore application profiles has been consistently informed by recognition of the importance of existing standards and best practices. In its metadata guidelines document, CanCore has utilized every available opportunity to reference established and emerging practices as a way of grounding its normative interpretations. Both TLF and CanCore further recognize that within learning technology standards communities, much effort has been expended on the development of bindings and schemas for the purposes of syntactic and systems-level interoperability, but that less attention has been paid to issues of semantic interoperability. Both TLF and CanCore recognize that this is not universally the case and that there is plenty of excellent work either already done or underway associated with semantics. For example, as illustrated above, CanCore utilizes definitions and explications found in Dublin Core itself and in work products of the broader DC community.

It is understood by both TLF and CanCore that interoperability – semantic or otherwise – is won by degrees, and often as a result of pragmatic efforts. It seems there will inevitably be a wide diversity in the communities of practice adopting metadata for application in learning, education, and training. However, it is our experience that pragmatic and open solutions are key to facilitating adoption.

### 7.2. Interoperability and Pragmatism

While "interoperability" seems to be a shared aim of any number of e-learning projects worldwide, it is clear that achieving it happens incrementally and often as a result of very deliberate and pragmatic efforts. Ultimately, there is a wide diversity in the communities of practice adopting metadata for application in learning, education, and training and it is our experience that pragmatic solutions are key to facilitating adoption.

## 8. Conclusion

Stable data models, combined with clearly delineated metadata community principles and practicalities, have facilitated development and implementation of both The Le@rning Federation and CanCore metadata application profiles. The experience of developing these two profiles has underscored the importance of identifying and responding to local requirements while at the same time respecting broader interoperability requirements. Of course, the true effectiveness of these application profiles will be tested when mechanisms for sharing or exchanging learning resources are put in place. It seems likely that further refinement of and reference between The Le@rning Federation metadata, CanCore, and other application profiles will be necessary in order for them to meet the needs of their stakeholders and of broader, cross-domain.interoperability requirements.

It is our shared view that continued and expanded dialogue on this topic would be greatly beneficial. In addition, learning resource metadata exchange testbeds and other test bed efforts would greatly enhance the interests of interoperability and resource sharing generally. Discussions regarding such collaboration between Australian and Canadian education authorities are already underway. It would be timely if similar efforts were undertaken across other domains and jurisdictions in the e-learning world. Although such work will no doubt presents daunting challenges, it is now urgently needed to realize the vision of interoperable and effective resource sharing.

## References

Australian Government, 2001. Backing Australia's Ability: Innovation Action Plan
http://backingaus.innovation.gov.au/

Duval, E., Hodgins, W., Sutton, S., Weibel, S.L., 2002. Metadata Principles and Practicalities. D-Lib Magazine, 8 (4).
http://www.dlib.org/dlib/april02/weibel/04weibel.html

DCMES 1999. Dublin Core Metadata Element Set, Version 1.1: Reference Description
http://www.dublincore.org/documents/dces/

DCQ 2000. Dublin Core Qualifiers
http://www.dublincore.org/documents/dcmes-qualifiers/

EdNA 2000. EdNA Metadata Homepage
http://standards.edna.edu.au/metadata/

Fisher, S, Friesen, N, Roberts, A. 2002. CanCore Intititive Metadata Guidelines Version 1.1
http://www.cancore.org/documents/

Friesen, N, Roberts, A. Fisher, S. 2002. CanCore: Learning Object Metadata
http://www.cancore.org/cancorepaper.doc

Heery, R., Patel, M., 2000. Application Profiles: Mixing and Matching Metadata Schemas. Ariadne, 25.

IEEE LOM, 2002. Learning Object Metadata draft version 6.4
http://ltsc.ieee.org/wg12/index.html

IMS CP, 2002.
http://www.imsglobal.org

ISO 1179. Specification and Standardization of Data Elements, Parts 1-6.
ftp://sdct-sunsrv1.ncsl.nist.gov/x3l8/11179/

Lynch, C.A. 1997. The Z39.50 Information Retrieval Standard. Part I: A Strategic View of Its Past, Present and Future. D-Lib Magazine. April. [Web Page].
http://www.dlib.org/dlib/april97/04lynch.html

ODRL 2002. Open Digital Rights Language
http://www.odrl.net

Ottawa 2001. The Ottawa Communique
http://www.ischool.washington.edu/sasutton/dc-ed/Ottawa-Communique.rtf

TLF 2002. The Le@rning Federation Metadata Application Profile
http://www.thelearningfederation.edu.au/repo/cms2/published/3059/metadata_application_profile_1.1.pdf

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 71-80*


# Exposing Cross-Domain Resources for Researchers and Learners

Ann Apps, Ross MacIntyre, Leigh Morris
MIMAS, Manchester Computing, University of Manchester,
Oxford Road, Manchester, M13 9PL, UK
ann.apps@man.ac.uk, ross.macintyre@man.ac.uk, leigh.morris@man.ac.uk

## Abstract

*MIMAS is a national UK data centre which provides networked access to resources to support learning and research across a wide range of disciplines. There was no consistent way of discovering information within this cross-domain, heterogeneous collection of resources, some of which are access restricted. Further these resources must provide the interoperable interfaces required within the UK higher and further education 'information environment'. To address both of these problems, consistent, high quality metadata records for the MIMAS services and collections have been created, based on Dublin Core, XML and standard classification schemes. The XML metadata repository, or 'metadatabase', provides World Wide Web, Z39.50 and Open Archives Initiative interfaces. In addition, a collection level database has been created with records based on the RSLP Collection Level Description schema. The MIMAS Metadatabase, which is freely available, provides a single point of access into the disparate, cross-domain MIMAS datasets and services.*
**Keywords.** *Metadata, Dublin Core, cross-domain, collection level description, subject classification.*

## 1. Introduction

MIMAS [26] at the University of Manchester, UK, is a national data centre for higher and further education and the research community in the UK, providing networked access to key data and information resources to support teaching, learning and research across a wide range of disciplines. This cross-domain, heterogeneous collection of resources includes:

- Bibliographic information such as ISI Web of Science, COPAC providing access to the UK research libraries' online catalogue, and the *zetoc* current awareness service based on the British Library's electronic table of contents database of journal articles and conference papers.
- Electronic journals via the JSTOR archive of scholarly journals, and the UK National Electronic Site Licence Initiative (NESLI).
- Archival information from the Archives Hub national gateway to descriptions of archives in UK universities and colleges.
- Statistical datasets including data from several UK censuses, international macro-economic data and UK government surveys.
- Spatial, satellite and geographic datasets.
- Scientific, chemical data via Beilstein Crossfire.
- Software packages for manipulating some of these datasets.

Until now there was no consistent way of discovering information within these MIMAS collections and associated services, except by reading the web pages specific to each service. Although most of these web pages contain high quality information relevant to their particular service, this information is not presented in a standard format and there is not a simple way to search for information across the services.

Some of the resources held at MIMAS are freely available globally, but access to many is restricted in some cases to members of UK academia, maybe requiring registration, in other cases by subscription. For resources where access is restricted, currently general resource discovery will find only shallow top-level information, and may not indicate to a prospective user the appropriateness of a resource to their interest.

MIMAS services are funded by the Joint Information Systems Committee (JISC) [17] of the UK Higher and Further Education Funding Councils. Thus they will be required to provide interfaces consistent with the architecture of the JISC 'Information Environment' [33] for resource discovery by researchers and learners. Currently many of the services do not provide these interfaces. Some of the MIMAS services are products hosted and supported by MIMAS, but not developed in-house, making implementation of additional interfaces unlikely.

To overcome all of these problems consistent, high quality metadata records for the MIMAS services and collections have been created. These metadata records are standards-based, using Dublin Core [7], XML and standard encoding schemes for appropriate fields. Freely available access to this XML metadata repository, or 'metadatabase', is provided by an application which supports the interfaces required by the Information Environment, enabling information discovery across the cross-domain MIMAS resource collection and allowing users at all experience levels access to 'an interoperable information world' [3].

## 2. MIMAS Metadata Records

### 2.1. Cross-Domain Information Discovery

Because the MIMAS service consists of a heterogeneous collection of services and datasets across many disciplines, a common, cross-domain metadata schema is required for their description. The metadata created to describe them is based on qualified Dublin Core, which enables cross-searching using a common core of metadata. This allows someone searching for information about for example 'economic' to discover results of possible interest across many of the MIMAS services beyond the obvious macro-economic datasets, including JSTOR, census data, satellite images and bibliographic resources. It is possible that in the future the metadata will be extended to include records according to domain-specific standards, such as the Data Documentation Initiative (DDI) Codebook [10] for statistical datasets or a standard geographic scheme, such as ISO DIS 19115 Geographic Information – Metadata [15], for census and map datasets. Another possible future extension would be to include educational metadata, such as IMS [30], where appropriate datasets are learning resources. But the MIMAS metadata cross searching capability would of necessity still be based on the 'core' metadata encoded in qualified Dublin Core.

### 2.2. An Example Metadata Record

The MIMAS metadata is encoded in XML and stored in a Cheshire II [19] database, described in more detail in section 5, which provides a World Wide Web and a Z39.50 interface. NISO Z39.50 [28] is a standard for information retrieval which defines a protocol for two computers to communicate and share information [25].

Using the Web interface to this metadatabase, searches may be made by fields *title*, *subject* or '*all*', initially retrieving a list of brief results with links to individual full records.

Following a Z39.50 search, records may be retrieved as Simple Unstructured Text Record Syntax (SUTRS), both brief and full records, full records being similar to the above example, GRS-1 (Generic Record Syntax) [23] and a simple tagged reference format. In addition the MIMAS Metadatabase is compliant with the Bath Profile [2], an international Z39.50 specification for library applications and resource discovery, providing records as simple Dublin Core in XML according to the CIMI Document Type Definition [5].

The MIMAS Metadatabase has the capability to expose simple Dublin Core metadata about the MIMAS resources for harvesting, conforming to the Open Archives Initiative (OAI) [29] Metadata Harvesting Protocol.

An example of a full record for one of the results retrieved by searching for a subject 'science', with web links underlined, but with an abbreviated description, is:

| | |
|---|---|
| Title: | ISI Web of Science |
| Creator: | MIMAS; ISI |
| Subject: (LCSH) | Abstracts; Arts; Books Reviews; Humanities; Letters; Periodicals; Reviews; Science; Social sciences |
| Subject: (UNESCO) | Abstracts; Arts; Book reviews; Conference papers; Discussions (teaching method Periodicals; Science; Social sciences |
| Subject (Dewey): | 300; 500; 505; 600; 605; 700; 705 |
| Description: | ISI Citation Databases are multidisciplinary databases of bibliographic information gathered from thousands of scholarly journals |
| Publisher: | MIMAS, University of Manchester |
| Type (DC): | Service |
| Type (LCSH): | Bibliographical citations; Bibliographical services; Citation indexes; Information retrieval; Online bibliographic searching; Periodicals Bibliography; Web databases |
| Type: (UNESCO) | Bibliographic databases; Bibliographic services; Indexes; Information retrieval; Online searching |
| Type (Dewey): | 005 |
| Type (MIMAS): | bibliographic reference |
| Medium: | text/html |
| URL: | http://wos.mimas.ac.uk/ |
| Language: | eng |
| isPartOf: | ISI Web of Science for UK Education |
| hasPart: | Science Citation Index Expanded |
| hasPart: | Social Sciences Citation Index |
| hasPart: | Arts & Humanities Citation Index |
| Access: | Available to UK FE, HE and research councils. Institutional subscription required |
| MIMAS ID: | wo000002 |

### 2.3. Standard Classification and Encoding Schemes

To provide quality metadata for discovery, subject keywords within the metadata are encoded according to standard classification or encoding schemes. These subject keywords will enable discovery beyond simply the existence of a resource by allowing researchers to find resources which are relevant to their particular research field. In order to facilitate improved cross-domain searching by both humans and applications where choices of preferred subject scheme might vary, MIMAS Metadata provides subjects encoded according to several schemes. As well as the encoding schemes currently recognised within qualified Dublin Core, Library of Congress Subject Headings (LCSH) [22] and Dewey Decimal [9], UNESCO [38] subject keywords are also available. In addition, MIMAS-specific subjects are included to capture existing subject keywords on the MIMAS web site service information pages supplied by the content or application creators as well as MIMAS support staff.

The use of standard classification schemes will improve resource discovery [40]. If schemes such as Dewey Decimal [37] were used, in the future, in a multi-faceted form they would lend themselves to use by developing search engines which create their indexes on faceted subject headings [11]. The development of more sophisticated ontology-based search engines will make the use of standard schemes even more important. Employing standard schemes will also assist in the provision of browsing structures for subject-based information gateways [18].

Similar classification schemes are included for 'Type' to better classify the type of the resource for cross-domain searching. Each metadata record includes a 'Type' from the high-level DCMI Type Vocabulary [8], 'Service' in the example above, but for some MIMAS records this will be 'Collection' or 'Dataset'. In addition, the above example includes type indications, including 'Bibliographical citations' and 'Online searching', according to standard schemes. Again the MIMAS-specific resource type is included.

Countries covered by information within a MIMAS service are detailed according to their ISO3166 [12] names and also their UNESCO names, captured within the 'dcterms:spatial' element of the metadata record and shown on the web display as 'Country'. This is of particular relevance to the macro-economic datasets, such as the IMF databanks, which include data from many countries in the world. Temporal coverage, again of relevance to the macro-economic datasets, is captured within a 'dcterms:temporal' element and encoded according to the W3CDTF [41] scheme. This is displayed as 'Time' and may consist of several temporal ranges. Information about access requirements to a particular MIMAS service is recorded as free-text within a 'dc:rights' element and displayed as 'Access'.

### 2.4. The MIMAS Application Profile

Where possible the metadata conforms to standard qualified Dublin Core. But this is extended for some Dublin Core elements to enable the capture of information which is MIMAS-specific or according to schemes which are not currently endorsed by Dublin Core. These local additions to qualified Dublin Core effectively make up the MIMAS application profile [14] for the metadatabase. The inclusion of UNESCO as a subject, type and spatial classification scheme described above is an example of local extensions, as is the capture of MIMAS-specific subjects and types. A possible future extension would be to capture the provenance of some metadata elements, such as subject keywords, where these were supplied by the content creator.

Some administrative metadata is included: the name of the person who created the metadata; the creation date; and the identifier of the record within the MIMAS Metadatabase. Capturing the name of the metadata creator will be of use for future quality checks and updating. The creation date, or 'date stamp', for the metadata, actually the date it is added into the database, is captured within a 'dcterms:created' element according to the W3CDTF scheme, for example "2002-05-27". The local MIMAS identifier, which is required to implement the functionality of the application as well as providing a unique identifier for each record within the database, is captured in a dc:identifier element with a MIMAS scheme.

### 2.5. The MIMAS Metadata Hierarchy

Although each of the records within the MIMAS Metadatabase is created, indexed and available for discovery individually, the records represent parts of the service within a hierarchy. In the example above, the record for 'ISI Web of Science' is a 'child' of the top-level record 'ISI Web of Science for UK Education', the umbrella term for the total service offered, and is a 'parent' of several records including 'Science Citation Index Expanded'.

During metadata creation only the 'isPartOf' relation is recorded, as the MIMAS identifier of the parent metadata record. The 'hasPart' fields and the displayed titles and links for parent and child metadata records are included by the MIMAS Metadatabase application as described in section 5.2. Hard coding 'hasPart' fields into a metadata record would necessitate the inefficient process of updating a parent record whenever a new child record were added. Dynamic generation of these links assists in simplifying the metadata creation and update process, and in maintaining the consistency of the metadata.

A further navigation hierarchy is provided by the application. If a parent and a child record, according to the 'isPartOf' hierarchy, also have a matching MIMAS subject keyword, the application includes a link from the parent's subject keyword to the particu-

lar child record. For example a JSTOR fragment record could include:

| | |
|---|---|
| Title: | JSTOR Ecology & Botany Collection |
| Subject (MIMAS): | Ecology / Journal of Applied Ecology |
| Subject (MIMAS): | Botany |

where the text 'Ecology / Journal of Applied Ecology' is a web link to the record for that particular journal. Again this subject navigation hierarchy is provided dynamically by the application and does not depend on the accuracy of metadata creation beyond the 'isPartOf' identifier and the matching subject keyword.

The child, 'hasPart', links within the MIMAS metadata hierarchy are available in the web interface only. A metadata record retrieved through the Z39.50 or OAI interfaces will include a single 'isPartOf' relation at most, which will consist of the MIMAS identifier of the parent record. Any required linking between records would be provided by the application retrieving the records.

### 2.6. Metadata Creation

The initial MIMAS metadata covering all the MIMAS services has been created by one person as part of the set-up project, much of it being scraped from the existing MIMAS service web pages. The metadata records for each service have been checked manually by the particular support staff, thus ensuring quality metadata for each MIMAS service. It is envisaged that the metadata will be maintained by the service support staff in the future, as part of the standard support process for each MIMAS service. There are currently 57 records in the metadatabase, distributed unevenly across 14 services (maximum 14, minimum 1) but this will increase when the metadata is extended to lower levels in the hierarchy.

The metadata reaches appropriate levels of the hierarchy, differing for each service, but it may be extended to greater depth in the future, possibly to the data level in some cases. For instance, the individual journals and issues included in JSTOR could be listed in the metadatabase.

Lacking a suitable XML authoring tool, the MIMAS metadata is currently created as XML files using an XML template and a text editor. The created XML is validated by parsing against an XML Document Type Definition before the record is indexed in the metadatabase. It is planned to develop a specific web-form tool for metadata creation and updating. This tool will capture metadata by field and include links to standard schemes for subject keyword selection and classification, the required XML being created at its back end, effectively transparently. The tool will be 'wiki style' [21] allowing a metadata creator to immediately 'publish' and view the eventual display of the record within the application before making a final 'commit' to the metadatabase. Such a tool will become essential when the metadata maintenance is performed by more than one person.

## 3. The JISC Information Environment

All MIMAS resources are part of the JISC 'Information Environment' [33], which provides resources for learning, teaching and research to UK higher and further education, and thus must be consistent with its architecture. The Information Environment will enable resource discovery through the various portals in its 'presentation layer', including the discipline specific UK Resource Discovery Network (RDN) hubs [35], the RDN also being one of the participating gateways in the European Renardus service [36]. Content providers in the 'provision layer' are expected to disclose their metadata for searching, harvesting and by alerting. This means that all resources within the Information Environment should have a Web search interface and at least some of the following for machine-to-machine resource discovery: a Z39.50 (Bath Profile cross-domain compliant) search interface; an OAI (Open Archives Initiative) [29] interface for metadata harvesting; and an RDF Site Summary (RSS) [32] alert channel capability. In addition resources may support OpenURL [39] for article discovery and location, where appropriate. This initiative, based on standard metadata and methods, may be seen as moving the UK academic information environment 'from isolated digital collections to an interoperable digital library' [3].

The majority of MIMAS resources have a Web search interface to provide resource discovery within their particular service. A few MIMAS services, COPAC, *zetoc* and the Archives Hub, provide Z39.50 interfaces. Some services, being commercial products hosted by MIMAS, may never provide Z39.50 searching or OAI metadata. To overcome this lack of requisite interfaces for MIMAS content and access restrictions on some of the services, the MIMAS Metadatabase will act as an intermediate MIMAS service within the 'provision layer' of the Information Environment, functioning as the main resource discovery service for MIMAS content.

The MIMAS Metadatabase does not currently include an RSS alert facility. If thought necessary within the Information Environment, it would be possible to include an alerting service in the future where appropriate, which could inform researchers when new datasets or journals were added to the MIMAS collection.

OpenURL support is not included because the metadatabase is not primarily concerned with article discovery, although this is relevant to several of the MIMAS services. There is work underway to investigate the provision of OpenURL linking within *zetoc*, and ISI Web of Science provides OpenURL linking to users whose institution has an OpenURL resolver.

## 4. MIMAS Collection Description

A further requirement of the Information Environment is a 'collection description service' [43], to allow portals within the 'presentation layer' to determine which content providers may have resources of interest to their users. This will maintain machine-readable information about the various resource collections available to researchers and learners within the Information Environment. A portal will ascertain from a collection description that a particular content provider may have resources of interest to an end-user, before pointing the end-user to the content service.

MIMAS has developed a further metadata application, implemented using the same architecture as the metadatabase, to provide collection description metadata for its resources, based on the Research Support Libraries Programme (RSLP) Collection Level Description (CLD) Schema [34]. The MIMAS Collection database contains a record for each top-level collection at MIMAS, corresponding to the top-level descriptions of the MIMAS services in the metadatabase, with Web, Z39.50 and OAI interfaces.

Similar to the metadatabase, standard schemes are used to provide quality concepts for collection discovery. It is probable that the common subject classification used within the Information Environment will be Dewey Decimal, but LCSH and UNESCO concepts are also provided to allow searching by other sources.

MIMAS has extended the RSLP CLD schema to include administrative metadata needed for date stamping of records and quality control, including the record creation date, the name of the metadata record creator and the local identifier for the record.

In the web interface, there is a 'Describes' field which is a web link to the corresponding top-level service record in the MIMAS Metadatabase application. This link is inserted automatically by the application, based on the local MIMAS identifier within the collection record, rather than being hard-coded by the metadata creator, thus avoiding maintenance problems. Following this link enables navigation to lower level records within the MIMAS Metadatabase hierarchy. Including this link between the two applications, and so effectively between the two databases, removes the necessity to replicate within the MIMAS Collection Description all the MIMAS service descriptions at lower levels in the hierarchy. It is intended that the MIMAS Collection database will remain an exclusively top-level description.

### 4.1. An Example MIMAS Collection Record

An example collection description for a MIMAS service, *zetoc*, is as follows (with some abbreviation):

| | |
|---|---|
| Collection Name: | zetoc |
| Concept (LCSH): | Arts; Business; Conference proceedings; Diseases; Economics; Engineering; Finance; Geography; History; Humanities; Language; Law; Library materials; Literature; Medical sciences; Medicine; Online library catalogs; Periodicals; Philosophy; Political science; Psychology; Religion; Science; Social sciences; Technology |
| (UNESCO): | Concept: Conference papers; Diseases Economics; Engineering; Finance; Law; Medical sciences; Periodicals; Science; Social sciences; Technology |
| Concept: (Dewey) | 050; 100; 105; 200; 300; 320; 330; 340; 400; 405; 500; 505; 600; 605; 610; 620; 700; 705; 800; 805; 900; 905 |
| Temporal Cover: | 1993/ |
| Description: | zetoc provides Z39.50-compliant access to the British Library's Electronic Table of Contents (ETOC) |
| Collection URL: | http://zetoc.mimas.ac.uk |
| Type (CLDT): | Catalogue.Library.Text |
| Accumulation: | 2000/ |
| Contents Date: | 1993/ |
| Accrual: | The database is updated nightly (active, deposit, periodic) |
| Legal Status: | Please see the Terms and Conditions of Use for further details |
| Access: | Available conditionally free to UK FE and HE. Available by institutional subscription to UK research councils, English NHS regions, CHEST associated and affiliated sites, and academic institutions in Ireland |
| Collector: | The British Library |
| Owner: | The British Library |
| Location: | Manchester Computing |
| Location URL: | http://zetoc.mimas.ac.uk |
| Administrator: | MIMAS |
| Admin Role: | Service provider |
| Admin Email: | info@mimas.ac.uk |
| Describes: | ze000001 |

### 4.2. Using the RSLP Collection Level Description Schema for Digital Collections

Because the RSLP schema was developed for the purpose of recording collections held by libraries and museums, some issues have arisen when using it to describe digital collections. Mostly these questions related to the irrelevance and apparent repetition of some of the fields, for instance the collection URL and the location URL in the above example. In many cases the distinction between 'collector' and 'owner' is not obvious. 'Physical location' is probably not of great importance for a digital collection which could easily be moved or distributed, and it is unlikely to be of interest to an end-user, whereas the physical location of a museum collection would be of significance. However, it is recognised that all the fields in the schema

are optional. In general the RSLP CLD, although not an 'official' standard, seems to provide a suitable common format for interoperable collection descriptions.

The application's Z39.50 interface provides, amongst other formats, simple Dublin Core in XML, for Bath Profile cross-domain compliancy and for interoperability within the JISC Information Environment. Similarly the OAI interface provides metadata records in simple Dublin Core. The mapping from the RSLP CLD to simple Dublin Core inevitably 'dumbs down' the information provided and loses some of the richness of the RSLP CLD schema. The Z39.50 SUTRS full record results, which are similar to the web display, maintain the full RSLP CLD information, but may not be very easily parsable. Thus it appears that to use these collection description records for machine-to-machine data interoperability within the JISC Information Environment a further metadata schema based on RSLP CLD will be required for OAI harvesting. Similarly such a schema could be incorporated into the results returned according to the Z39.50 standard if a new profile were agreed.

## 5. The Cheshire II Information Retrieval System

The software platform used for the MIMAS Metadatabase is Cheshire II [20] which is a next generation online catalogue and full text information retrieval system, developed using advanced information retrieval techniques. It is open source software, free for non-commercial uses, and operates with open-standard formats such as XML and Z39.50, all reasons which influenced its choice for this project. Cheshire II was developed at the University of California-Berkeley School of Information Management and Systems, underwritten by a grant from the US Department of Education. Its continued development by the Universities of Berkeley and Liverpool receives funding from the Joint Information Systems Committee (JISC) of the UK Higher and Further Education Funding Councils and the US National Science Foundation (NSF). Experience and requirements from the development of the MIMAS Metadatabase have been fed back into the continuing Cheshire development. Although using evolving software has caused some technical problems, the Cheshire development team has been very responsive to providing new functionality, and this relationship has proved beneficial to both projects. Examples of new functionality are the sorting of result sets within the Cheshire Web interface and 'virtual' databases, described further in [1].

### 5.1. Z39.50 via Cheshire

Cheshire provides indexing and searching of XML (or SGML) data according to an XML Document Type Definition (DTD), and a Z39.50 interface. The underlying database for the MIMAS Metadatabase is a single XML data file containing all the metadata records, along with a set of indexes onto the data. The MIMAS metadata XML is mapped to the Z39.50 Bib-1 Attribute Set [4] for indexing and searching. The application's Z39.50 search results formats are detailed above in section 2.2. The mapping from the MIMAS metadata to the GRS-1 Tagset-G [23] elements is defined in the Cheshire configuration file for the database and is used by Cheshire to return data in GRS-1 format to a requesting client. The other Z39.50 result formats are implemented by bespoke filter programs which transform the raw XML records returned by Cheshire, the 'hooks' to trigger these filters being specified in the configuration file for the database. The mapping from the MIMAS metadata to simple Dublin Core, as required by the Bath Profile, is straightforward, the base data being qualified Dublin Core, albeit with some loss of information such as subject schemes. In order to obviate this information loss as much as possible, such details are included in parentheses in the supplied record. For example, a Z39.50 XML result for the example in section 2.2 may contain the element:

```
<subject>(LCSH) Abstracts</subject>
```

### 5.2. The Cheshire Web Interface

Cheshire also provides 'webcheshire' which is a basic, customisable World Wide Web interface. The web interface for the MIMAS Metadatabase is built on webcheshire as a bespoke program written in OmniMark (version 5.5) [31]. This web program provides a search interface which includes saving session information between web page accesses. It transforms retrieved records from XML to XHTML (version 1.0) for web display. OmniMark was chosen as the programming language for this interface because it is XML (or SGML) aware according to a DTD, a knowledge which is employed for the XML translations involved, and also because of existing expertise and availability on the MIMAS machine. Other suitable languages for the web interface implementation would have been Perl, or TCL which is the basic interface language to Cheshire.

The MIMAS Metadatabase web interface provides search results in discrete 'chunks', currently 25 at a time, with 'next' and 'previous' navigation buttons. This is implemented by using the Cheshire capability to request a fixed number of records in the result set, beginning at a particular number within that set. The application remembers the MIMAS identifiers of the results in the retrieved 'chunk', and extracts the record corresponding to a particular MIMAS identifier when an end-user selects a 'full record display'.

To implement the metadata hierarchy navigation functionality, described in section 2.5, an additional index, used internally by the application, is created

on the 'isPartOf' fields of the records which denote the MIMAS identifiers of the parent records. When a record is displayed, this index is checked to find all metadata records which indicate the current record as parent, the titles of these children records also being determined from the database. For each child record found a 'hasPart' link is displayed. Similarly the title and link for the 'isPartOf' display are determined by a database look-up.

Within the MIMAS Collection database, when a record is displayed, the MIMAS Metadatabase is checked for a record with a matching identifier. If such a record is found the display includes a 'Describes' web link from the Collection database to the corresponding record in the metadatabase.

## 6. Exposing OAI Metadata

The Open Archives Initiative (OAI) has specified a Metadata Harvesting Protocol [42] which enables a data repository to expose metadata about its content in an interoperable way. The architecture of the JISC Information Environment includes the implementation of OAI harvesters which will gather metadata from the various collections within the Information Environment to provide searchable metadata for portals and hence for end-users [6]. Portals will select metadata from particular subject areas of relevance to their user community. Thus there is a requirement for collections and services within the Information Environment to make their metadata available according to the OAI protocol, including a minimum of OAI 'common metadata format', i.e. simple Dublin Core, records.

An OAI metadata harvesting interface has been added to both the MIMAS Metadatabase and the Collection database, as a 'cgi' program, written in TCL which is the native language of Cheshire. This program responds appropriately to OAI requests, implementing the OAI 'verbs': *Identify* which details the database and its OAI level of support; *ListMetadataFormats* to indicate the metadata formats available, currently only simple Dublin Core (oai_dc); *ListIdentifiers* to list the identifiers of all the available records; *GetRecord* to retrieve the metadata of a particular record; *ListRecords* to list the metadata of all records; and *ListSets* which returns an empty response, sets not being supported.

In order to implement the OAI interface, three new search result formats have been defined for the databases, which return in XML, respectively, according to the required OAI format: the identifier of a record; the metadata of the record in Dublin Core; an identifier and date stamp for a record, where an unavailable metadata format is requested. The OAI cgi program performs the search on the Cheshire database according to the appropriate result format for the OAI verb and arguments, then passes the result to the harvester wrapped by the required OAI response format.

### 6.1. An Example OAI Record

An example response to a GetRecord request would be as follows, abbreviated for conciseness:

```
<?xml version="1.0" encoding="UTF-8" ?>
<GetRecord
xmlns=
"http://www.openarchives.org/OAI/1.1/OAI_GetRecord"
xmlns:xsi=
"http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation=
"http://www.openarchives.org/OAI/1.1/OAI_GetRecord
http://www.openarchives.org/OAI/1.1/OAI_GetRecord.xsd">
<responseDate>
 2002-05-28T11:59:45+01:00
</responseDate>
<requestURL>
 http://irwell.mimas.ac.uk/cgi- bin/cgiwrap/zzmetadm/
 mimas_oai?
 verb=GetRecord&identifier=oai%3Amimas%3Aze000001
  &metadataPrefix=oai_dc
</requestURL>
<record>
 <header>
  <identifier>oai:mimas:ze000001</identifier>
  <datestamp>2002-04-24</datestamp>
 </header>
 <metadata>
  <dc xmlns="http://purl.org/dc/elements/1.1/"
      xmlns:xsi=
        "http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="http://purl.org/dc/elements/1.1/
        http://www.openarchives.org/OAI/1.1/dc.xsd">
   <title>zetoc</title>
   <creator>The British Library</creator>
   <creator>MIMAS</creator>
   <subject>(Dewey) 050</subject>
   <identifier>http://zetoc.mimas.ac.uk</identifier>
  </dc>
 </metadata>
</record>
</GetRecord>
```

### 6.2. Date Range

The OAI protocol allows harvesters to specify they want records 'from' a certain date and/or 'until' a certain date. Selecting records added to the Cheshire database before or after a certain date, in response to an OAI request, is implemented easily when a Cheshire index has been created for the 'date loaded' (dcterms:created) field. This field is also used to provide the date stamp on returned records.

### 6.3. Response Restriction

With no restrictions, OAI harvesting could result in effective 'denial of service' attacks because of the machine resources required, so there is generally a

need for an 'acceptable use' policy to restrict how many records may be harvested at one time and how soon a repeat request may be made. This is probably not a serious consideration for the MIMAS Metadatabase, which currently does not contain a large number of records, but would be a significant issue if OAI interfaces were to be implemented on some of the very large datasets at MIMAS such as *zetoc* [1]. When supplying only part of a result set, the OAI protocol allows for the return of a 'resumptionToken' which the harvester uses to make repeat requests. The format of this 'resumptionToken' is not defined in the OAI protocol but by the source application.

The MIMAS Metadatabase returns a fixed number of records or identifiers in response to one request. If there are more records available a resumptionToken is returned. Because a repeat request will contain just the resumptionToken as an exclusive argument, details of the original request are included in the token to enable a repeat of the original search on the database.

The format of the resumptionToken for the MIMAS Metadatabase is:

<database>-<start>-<from>-<until>-<format>

where:

<database> is the database identifier
<start> is the number of the next record to be retrieved within the result set
<from> is the 'from' date specified in the original request (yyyymmdd) or zero
<until> is the 'until' date specified in the original request (yyyymmdd) or zero
<format> is the metadata format specified in the original request. This may be: 'dc' for Dublin Core; 'xx' for an unsupported metadata format; 'li' for a ListIdentifiers request where metadata format is irrelevant.

For example, a resumptionToken returned by a ListRecords request for Dublin Core records from 2002-04-01 until 2002-07-01 following the first 50 records would be:

mimas-51-20020401-20020701-dc

When an OAI request includes a resumptionToken, the cgi program parses the token, then performs the original search on the database, but requesting a result set beginning at the token's <start> number. For a large result set, this search may again result in a further resumptionToken. This implementation relies on Cheshire functionality which allows a search request to return a fixed number of results beginning at a stated point within the result set.

### 6.4. Subject Keywords in OAI Records

Because simple Dublin Core metadata format records are supplied to OAI harvesters, there is some loss of richness in the information from the base qualified Dublin Core data, similar to that described for Z39.50 XML results in section 5.1. In particular, the subject encoding scheme used is not included, unless in parentheses as part of a subject keyword text string. Knowledge of the encoding schemes used for subject keywords would be important to services which are providing search interfaces across metadata harvested from multiple repositories, both to ensure the quality of the metadata and for comparison between subject keywords from harvested sources [24]. If a qualified Dublin Core XML schema were available, and recognised by OAI and the JISC Information Environment, then more complete metadata, including relevant encoding schemes, could be supplied to metadata harvesters from the MIMAS Metadatabase.

## 7. Conclusion

MIMAS has aimed to describe its collection of datasets and services using quality metadata. Quality assurance has been achieved by checking of the metadata records for a particular service by the relevant support staff. Continued metadata quality will be ensured by maintenance of the metadata by these support staff. Subject or concept keywords are included in the metadata according to several standard classification schemes, as are resource types and geographical names. Use of standard schemes enhances the quality of the metadata and enables effective resource discovery.

Another objective of the project was to develop an interoperable solution based on open standards and using leading-edge, open source technology. This has been successfully achieved using a Cheshire II software platform to index Dublin Core records encoded in XML. A spin-off has been improvements to Cheshire following feedback from MIMAS. Use of other standard or experimental technologies such as the Z39.50 and OAI metadata harvesting interfaces in addition to the web interface will enable the MIMAS Metadatabase and Collection database to be integrated into the JISC 'Information Environment', thus providing a valuable resource discovery tool to the stakeholders within that environment.

The MIMAS Metadatabase provides a single point of access into the disparate, cross-domain MIMAS datasets and services. It provides a means for researchers to find and access material to aid in the furtherance of their work, thus assisting in the advancement of knowledge. Learners and their teachers will be able to discover appropriate learning resources across the MIMAS portfolio, improving the educational value of these datasets.

The MIMAS Metadatabase may be searched at http://www.mimas.ac.uk/metadata/ and the MIMAS Collection Description at http://www.mimas.ac.uk/metadata/collection/ .

## Acknowledgements

The authors wish to acknowledge the contribution to the development of the MIMAS Metadatabase by their colleagues at MIMAS who support the many services and datasets and provided quality metadata, and the Cheshire development team, Ray Larson at the University of California-Berkeley and Paul Watry and Robert Sanderson at the University of Liverpool. The development of the MIMAS Metadatabase and the MIMAS Collection database was funded as part of the 'Implementing the DNER Technical Architecture at MIMAS' (ITAM) project [16] and its predecessor 'MIMAS Metadata for the DNER' [27] by the Joint Information Systems Committee (JISC) for the UK Higher and Further Funding Councils within the Distributed National Electronic Resource (DNER) programme [13].

## References

[1] Apps, A. and MacIntyre, R., "Prototyping Digital Library Technologies in zetoc", *Proceedings of ECDL2002: 6th European Conference on Research and Advanced Technology for Digital Libraries, Rome, September 16-18, 2002, 2002*, accepted for publication.

[2] The Bath Profile: An International Z39.50 Specification for Library Applications and Resource Discovery. http://www.nlc-bnc.ca/bath/bp-current.htm

[3] Besser, H., "The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries", *First Monday, 7 (6)*, 2002. http://firstmonday.org/issues/issue7_6/besser/

[4] Bib-1 Attribute Set. http://lcweb.loc.gov/z3950/agency/defns/bib1.htm

[5] The Consortium for the Computer Interchange of Museum Information (CIMI) Dublin Core Document Type Definition.
http://www.nlc-bnc.ca/bath/bp-app-d.htm

[6] Cliff, P., "Building ResourceFinder", *Ariadne*, 30, 2001. http://www.ariadne.ac.uk/issue30/rdn-oai/

[7] The Dublin Core Metadata Initiative. http://www.dublincore.org

[8] DCMI Type Vocabulary. http://dublincore.org/documents/dcmi-type-vocabulary/

[9] Dewey Decimal Classification. OCLC Forest Press. http://www.oclc.org/dewey/

[10] DDI, Data Documentation Initiative Codebook DTD. http://www.icpsr.umich.edu/DDI/CODEBOOK/

[11] Devadason, F., Intaraksa, N., Patamawongjariya, P. and Desai, K., "Search interface design using faceted indexing for Web resources", *Proceedings of the 64th ASIST Annual Meeting*. Medford: Information Today Inc, 38, 2001, pp. 224-238.

[12] ISO 3166-1: The Code List. http://www.din.de/gremien/nas/nabd/iso3166ma/codlstp1/

[13] The UK Distributed National Electronic Resource. http://www.jisc.ac.uk/dner/

[14] Heery, R. and Patel, M., "Application profiles: mixing and matching metadata schemas", *Ariadne*, 25, 2000.
http://www.ariadne.ac.uk/issue25/app-profiles

[15] ISO/TC211: Geographic Information/Geomatics. http://www.isotc211.org

[16] Implementing the DNER Technical Architecture at MIMAS (ITAM) project. http://epub.mimas.ac.uk/itam.html

[17] JISC, The Joint Information Systems Committee. http://www.jisc.ac.uk

[18] Koch, T. and Day, M., "The role of classification schemes in Internet resource description and discovery", *Work Package 3 of Telematics for Research project DESIRE (RE 1004)*, 1999. http://www.ukoln.ac.uk/metadata/desire/classification/

[19] Larson, R.R., Cheshire II Project. http://cheshire.lib.berkeley.edu

[20] Larson, R.R., McDonough, J., O'Leary, P., Kuntz, L., Moon, R., "Cheshire II: Designing a Next-Generation Online Catalog", *Journal of the American Society for Information Science*, 47 (7), 1996, pp. 555-567.

[21] Leuf, B. and Cunningham, W., The Wiki Way. http://www.wiki.org

[22] Library of Congress Subject Headings. *Cataloguing Distribution Service*, Library of Congress. http://lcweb.loc.gov/cds/lcsh.htm

[23] The Z39.50 Generic Record Syntax (GRS-1) Tagsets.
http://lcweb.loc.gov/z3950/agency/defns/tag-gm.html

[24] Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M.L., Knudson, F. and Holtkamp, I., "Federated Searching Interface Techniques for Heterogeneous OAI Repositories", *Journal of Digital Information*, 2 (4), 2002, http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/

[25] Miller, P., "Z39.50 for All", *Ariadne*, 21, 1999, http://www.ariadne.ac.uk/issue21/z3950

[26] MIMAS, Manchester Information and Associated Services, including Archives Hub, COPAC, ISI Web of Science, JSTOR, NESLI, zetoc. http://www.mimas.ac.uk

[27] MIMAS Metadata for the DNER project. http://epub.mimas.ac.uk/dner.html

[28] Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. http://www.niso.org/standards/resources/Z3950.pdf

[29] Open Archives Initiative. http://www.openarchives.org

[30] Olivier, B., Liber, O. and Lefrere, P., "Specifications and standards for learning technologies: the IMS project", *International Journal for Electrical Engineering Education*, 37 (1), 2000, pp. 26-37.

[31] OmniMark Technologies, http://www.omnimark.com

[32] Powell, A., "RSS FAQ, JISC Information Environment Architecture", 2002. http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/faq/rss/

[33] Powell, A. and Lyon, L., "The JISC Information Environment and Web Services", *Ariadne*, 31, 2002. http://www.ariadne.ac.uk/issue31/information-environments.

[34] Powell, A., Heaney, M. and Dempsey, L., "RSLP Collection Description", *D-Lib Magazine*, 6 (9), 2000. doi:/10/1045/september2000-powell.

[35] RDN, Resource Discovery Network. http://www.rdn.ac.uk

[36] Renardus. http://www.renardus.org

[37] Tinker, A.J., Pollitt, A.S., O'Brien, A. and Braekevelt, P.A., "The Dewey Decimal Classification and the transition from physical to electronic knowledge organisation", *Knowledge Organization*, 26 (2), 1999, pp. 80-96.

[38] UNESCO Thesaurus. http://www.ulcc.ac.uk/unesco/

[39] Van de Sompel, H., Beit-Arie, O., "Open Linking in the Scholarly Information Environment Using the OpenURL Framework", *D-Lib Magazine*, *7 (3)*, 2001. doi://10.1045/march2001-vandesompel

[40] Vizine-Goetz, D., "Using Library Classification Schemes for Internet Resources", E. Jul, ed. *Proceedings of the OCLC Internet Cataloguing Colloquium, San Antonio, Texas, 19 January 1996*. OCLC, 1996.
http://www.oclc.org/oclc/man/colloq/toc.htm

[41] W3C, Date and Time Formats. http://www.w3.org/TR/NOTE-datetime

[42] Warner, S., "Exposing and Harvesting Metadata Using the OAI Metadata Harvesting Protocol: A Tutorial", *High Energy Physics Webzine*, *4*, 2001, http://library.cern.ch/HEPLW/4/papers/3/

[43] Watry, P. and Hill, A., "Collection Description Service Scoping Study", JISC Report, 2002.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 81-89*


# Integrating Schema-specific Native XML Repositories into a RDF-based E-Learning P2P Network

Changtao Qu
Learning Lab Lower Saxony
University of Hannover
Expo Plaza 1, D-30539
Hannover, Germany
qu @learninglab.de

Wolfgang Nejdl
Computer Science Dept.
Stanford University, Stanford, CA 94305, USA
nejdl @db.stanford.edu
(On leave from University of Hannover)

Holger Schinzel
Learning Lab Lower Saxony
University of Hannover
Expo Plaza 1, D-30539
Hannover, Germany
schinzel @learninglab.de

## Abstract

*As its name implies, a native XML repository supports storage and management of XML in the original hierarchical form rather than in some other representations. In this paper we present our approach for integrating native XML repositories into Edutella, a RDF-based E-learning P2P network, through mapping native XML database schemas onto the Edutella Common Data Model (ECDM) and further translating ECDM's internal query language Datalog into XPath, the local query language of native XML repositories. Due to the considerable incomparability between the ECDM and the XML data model, a generic integration approach for schema-agnostic native XML repositories is found to be unrealistic. Thus our investigations are focused on three schema-specific native XML repositories respectively based on the DCMES, LOM/IMS, and SCORM XML binding data schema. Since these three metadata sets are the most popularly applied learning resource metadata specifications in E-Learning, our integration approach satisfactorily addresses the current usage of Edutella in E-Learning despite that a generic integration approach for schema-agnostic native XML repositories has not been implemented.*
**Keywords:** *repositories, E-learning network.*

## 1. Introduction

The open source project Edutella[1] is a RDF (Resource Description Framework)-based E-Learning P2P (Peer-to-Peer) network that aims at accommodating distributed learning resource metadata repositories, which are generally heterogonous in applied back-end systems, applied metadata schemas, etc., in a P2P manner and further facilitating the exchange of learning resource metadata between these repositories based on RDF [16]. At present Edutella is geared towards learning resource metadata repositories that are constructed based on three popular learning resource metadata sets: DCMES (Dublin Core Metadata Element Set)[7], IEEE LOM (Learning Object Metadata)/IMS Learning Resource Metadata Specification [11][12],

[1] http://edutella.jxta.org

and ADL (Advanced Distributed Learning) SCORM (Sharable Content Object Reference Model)[1], though its architecture and design does not make any assumptions about the applied metadata sets. In Edutella we make only one essential assumption that all Edutella resources can be described in RDF and further all Edutella functionalities can be mediated through RDF statements and the queries on these statements, as we believe the modular nature of RDF metadata to be especially suitable for distributed P2P settings. This essential assumption obviously leads to RDF being the most naturally applicable metadata representation in the Edutella network and thus RDF-based repositories containing the metadata of RDF bindings to above three learning resource metadata specifications are the most natural form of Edutella content provider peers.

However, in practice we currently have to address another important form of Edutella content providers: the XML (eXtensible Markup Language)-based repositories containing the metadata of XML bindings to the three learning resource metadata sets mentioned above. As a matter of fact, at present the XML-based learning resource metadata repositories still occupy a quite dominant place in E-Learning in comparison to the RDF-based repositories, although the latter ones have recently found more and more application cases [4][8]. Besides the reason that simple XML has a flatter learning curve and also a more straightforward binding strategy to all three learning resource metadata specifications in comparison to RDF, another important reason lies in the fact that XML has a longer history to be applied for binding learning resource metadata specifications than RDF. Taking the LOM/IMS metadata specification as an example, it has provided the XML binding since version 1.0, released in August 1999, whereas its RDF binding has only been introduced since version 1.2, released in June 2001. As a direct consequence, currently most of existing learning resource metadata repositories are XML-based [9][15][17][19], containing a large number of learning resource metadata to be addressed by Edutella.

In addition, the XML-based repositories also introduce a new type of back-end system: the native XML database, which provides a very straightforward way for constructing learning resource metadata repositories in that all learning resource XML metadata profiles can be directly stored and managed in the native XML repositories without the need of any pre-processing. The native XML databases support storage and management of XML in the original hierarchical form rather than in some other representations, e.g., decomposed relational tables in RDBs (Relational Database), or decomposed objects in OODBs (Object-oriented Database). Moreover, in a native XML database, the database schema used to define how the XML is stored is virtually identical to the XML data schema defined by XML DTD (Document Type Definition) or W3C (World Wide Web Consortium) XML Schema [20]. Based on a specific XML data schema, multiple XML metadata profiles can be contained in a single collection and thus can be queried as a whole through using W3C XPath [6], the query language supported by almost all native XML databases. Also the stored XML metadata profiles can be easily updated through direct manipulation on XML fragments instead of on the whole profiles. As a matter of fact, these features of the native XML databases satisfactorily fit into the typical usage and management scenarios of learning resource metadata and thus greatly promote the application of the native XML repositories in E-Learning.

However, despite of the fact that the XML-based learning resource metadata repositories have been popularly applied in E-Learning, there exists a big obstacle to integrate them into the RDF-based Edutella network. This obstacle comes from the considerable incomparability between RDF's binary relational data model and XML's hierarchical data model, which makes it difficult to establish the mapping from an arbitrary XML data schema to the RDF data model, although the reverse mapping is definitely feasible [14]. Therefore, in this paper we will mainly concentrate on three schema-specific native XML repositories, which accommodate learning resource metadata respectively based on the DCMES, LOM/IMS, and SCORM XML binding schema, and present our approach for integrating them into the RDF-based Edutella network. Since these three metadata sets are the most popularly applied learning resource metadata specifications in E-Learning, our integration approach satisfactorily addresses the current usage of Edutella in E-Learning despite that a generic integration approach for schema-agnostic native XML repositories has not been implemented.

## 2. Edutella provider integration architecture

Edutella employs a wrapper-like architecture for integrating heterogeneous content provider peers. In figure 1 we illustrate the Edutella provider integration architecture.

The wrapper-like architecture has been popularly applied for integrating heterogeneous information sources for many years [10][18]. The key to such sort of integration architecture is a common data model that is shared by all information sources and provides the common data view of the underlying heterogeneous repositories. For each wrapper program, it is on the one hand responsible for generating the common data view of the individual repository based on the pre-defined common data model, on the other hand, it is also responsible for translating the common query language for the common data view into the local query language of the individual repository, and vice versa, transforming the local query results into the results represented by the common result

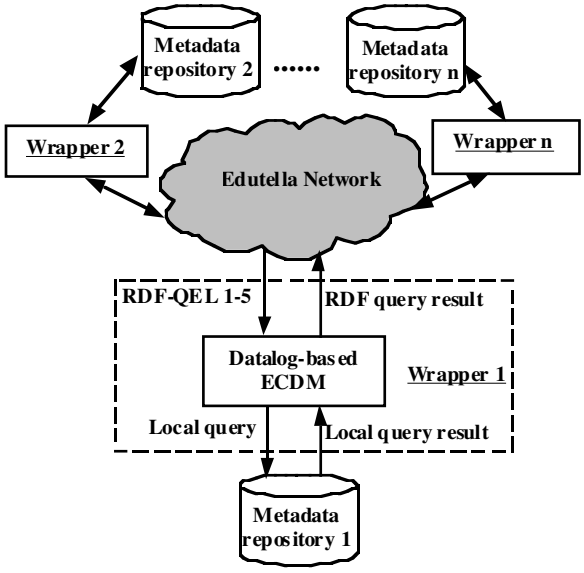



**Figure 1. The Edutella provider integration architecture**

exchange format after the query against the individual repository is completed.

Following this design and usage scenario, in Edutella we first proposed the Edutella Common Data Model (ECDM), which is defined in full compliance with the RDF data model and uses Datalog [10] as its internal query language. Externally, we defined a common query language: RDF Query Exchange Language (RDF-QEL) for the whole Edutella network using RDF syntax. As illustrated in Figure 1, the wrapper program of each Edutella provider is responsible for translating RDF-QEL into ECDM's internal query language Datalog. Because Datalog and RDF share the central feature that their relational data models are based on sets of ground assertions conceptually grouped around properties, there exists a natural approach for generating the ECDM-based common data view of the RDF-based repositories, as well as a natural approach for translating RDF-QEL into Datalog, which is internally used to manipulate the ECDM-based common data view.

Based on the wrapper-like Edutella provider integration architecture, we have successfully integrated several heterogeneous content provider peers into the Edutella network [16]. However, when we tried to handle the native XML repositories, several issues had to be addressed.

First, the XML data model is in some sense quite incomparable to the ECDM, which makes it difficult to integrate schema-agnostic XML-based repositories into Edutella. The ECDM, which is compliant with the RDF data model as well as the Datalog data model, is at its basis a binary relational data model consisting of a set of ground assertions represented either as binary predicates: *Predicate(Subject, Object)* or as ternary statements *s(Subject, Predicate, Object)*, if the predicate is taken as an additional argument. In contrast to the ECDM, the XML data model, which possesses a tree-like hierarchical data structure, cannot be easily mapped onto a binary relational data model, especially when the XML data schemas become complex enough, e.g., containing recursive elements, as it occurs in the LOM/IMS XML binding [12]. Moreover, in comparison to some powerful query languages supported by RDBs and OODBs, which can be used to generate the ECDM-based common data view of the underlying repositories, the XPath query language, which is currently the most used tool for manipulating the native XML repositories, is much weaker and thus incapable of manipulating some complex XML data models to generate their ECDM-based common data view. This incomparability between the XML data model and the ECDM influenced our decision to apply our integration approach only to several schema-specific XML repositories at the current time.

Second, in comparison to ECDM's internal query language Datalog, XPath is also far from comparable and thus cannot express all Datalog queries. Whereas Datalog is a relationally complete query language that is able to express relational algebra such as "selection", "union", "join", and "projection", etc., and also possesses some additional features such as transitive closure and recursive definitions, XPath can only express part of relational algebra, such as "union", limited "selection", and "negation" in terms of the XML tree-like data model, but lacks the support for expressing "join" and "projection". As introduced in our previous publication [16], at present we have defined five sets of RDF-QELs in the Edutella network according to their different expressivity, namely, RDF-QEL1 (can express conjunctive query), RDF-QEL2 (RDF-QEL1 plus disjunctive query), RDF-QEL3 (RDF-QEL2 plus query negation), RDF-QEL4 (RDF-QEL3 plus linear recursive query), and RDF-QEL5 (RDF-QEL4 plus arbitrary recursive query), all of which can be transparently translated into the corresponding Datalog queries. While all sets of RDF-QEL queries can be fully handled by some high-performance RDF-based repositories such as RDBs supporting SQL3, the native XML repositories can only handle part of the RDF-QEL sets, namely, RDF-QEL1 to RDF-QEL3. In fact, the weak expressivity of XPath determines that the native XML repositories in the Edutella network are unable to achieve the same functionalities as other high-performance repositories with the support of some powerful local query languages.

Finally, the incomparability between the XML data model and the ECDM as well as the incomparability between Datalog and XPath also have a negative influence on the query result representation of the native XML repositories. Whereas the RDF-based repositories can naturally adapt the query results into Edutella's RDF-based common result exchange format with the support of some powerful local

query languages, the native XML repositories can only return XML fragments selected by the XPath expressions rather than sets of tuples that can be naturally brought into the RDF model due to XPath's limited capability of expressing "selection", as well as its incapability of expressing "join" and "projection". Therefore, the query results generated by the native XML repositories need some additional processing in order to be adapted into the Edutella common result exchange format.

In the following we will present our approach addressing above issues. The native XML repository introduced here is implemented using the open source project Apache Xindice 1.0[2], but the presented approach is also applicable to some other native XML repository implementations, e.g., Tamino XML database 3.1.1.1[3], Ipedo XML database 3.0.1[4], etc. In addition, although our approach will address three schema-specific native XML repositories that accommodate learning resource metadata respectively based on the DCMES, LOM/IMS, and SCORM XML binding schema, we will use the DCMES, which constitutes the minimal interoperable basis of some more complicated metadata sets, as the "standard" schema throughout the discussion. In section 6 we will describe the integration approach, which is based on the DCMES XML binding data schema, for integrating the LOM/IMS and SCORM XML binding schema based native XML repositories into the Edutella network.

## 3. Generating the ECDM-based common data view of the native XML repositories

The DCMES XML binding [2] is the guideline proposed by DCMI (Dublin Core Metadata Initiative) for the XML encoding of DCMES. The primary goal of this guideline is to provide a simple DCMES encoding, where there are no extra elements, qualifiers, operational or varying parts allowed. The secondary goal is to make the encoding also be valid RDF, which allows the XML binding to be manipulated using the RDF model. For the DCMES XML binding schema based native XML repositories contained in the Edutella network, the second design goal of the DCMES XML binding to a certain degree facilitates the adaptation of their local query results into the Edutella common result exchange format[5].

In Figure 2 we show the XML schema of the DCMES XML binding in the format of XML DTD [2].

The above XML schema can be also viewed in a schematic way, represented in the hedgehog model, as depicted in Figure 3 [13].

From the hedgehog model of the DCMES XML binding, in which all assertions are made about a

```
<!ENTITY rdfns 'http://www.w3.org/1999/02/22-rdf-syntax-ns#' >
<!ENTITY dcns 'http://purl.org/dc/elements/1.1/' >
<!ENTITY % rdfnsdecl 'xmlns:rdf CDATA #FIXED "&rdfns;"' >
<!ENTITY % dcnsdecl 'xmlns:dc CDATA #FIXED "&dcns;"'>
<!ELEMENT rdf:RDF (rdf:Description)* >
<!ATTLIST rdf:RDF %rdfnsdecl; %dcnsdecl; >
<!ENTITY % dcmes "dc:title | dc:creator | dc:subject | dc:description|
dc:publisher | dc:contributor | dc:date | dc:type | dc:format | dc:identifier | dc:source | dc:language | dc:relation | dc:coverage |
dc:rights" >
<!ELEMENT rdf:Description (%dcmes;)* >
<!ATTLIST rdf:Description rdf:about CDATA #IMPLIED>
<!ELEMENT dc:title (#PCDATA)>
<!ATTLIST dc:title xml:lang CDATA #IMPLIED>
<!ATTLIST dc:title rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:creator (#PCDATA)>
<!ATTLIST dc:creator xml:lang CDATA #IMPLIED>
<!ATTLIST dc:creator rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:subject (#PCDATA)>
<!ATTLIST dc:subject xml:lang CDATA #IMPLIED>
<!ATTLIST dc:subject rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:description (#PCDATA)>
<!ATTLIST dc:description xml:lang CDATA #IMPLIED>
<!ATTLIST dc:description rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:publisher (#PCDATA)>
<!ATTLIST dc:publisher xml:lang CDATA #IMPLIED>
<!ATTLIST dc:publisher rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:contributor (#PCDATA)>
<!ATTLIST dc:contributor xml:lang CDATA #IMPLIED>
<!ATTLIST dc:contributor rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:date (#PCDATA)>
<!ATTLIST dc:date xml:lang CDATA #IMPLIED>
<!ATTLIST dc:date rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:type (#PCDATA)>
<!ATTLIST dc:type xml:lang CDATA #IMPLIED>
<!ATTLIST dc:type rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:format (#PCDATA)>
<!ATTLIST dc:format xml:lang CDATA #IMPLIED>
<!ATTLIST dc:format rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:identifier (#PCDATA)>
<!ATTLIST dc:identifier xml:lang CDATA #IMPLIED>
<!ATTLIST dc:identifier rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:source (#PCDATA)>
<!ATTLIST dc:source xml:lang CDATA #IMPLIED>
<!ATTLIST dc:source rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:language (#PCDATA)>
<!ATTLIST dc:language xml:lang CDATA #IMPLIED>
<!ATTLIST dc:language rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:relation (#PCDATA)>
<!ATTLIST dc:relation xml:lang CDATA #IMPLIED>
<!ATTLIST dc:relation rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:coverage (#PCDATA)>
<!ATTLIST dc:coverage xml:lang CDATA #IMPLIED>
<!ATTLIST dc:coverage rdf:resource CDATA #IMPLIED>
<!ELEMENT dc:rights (#PCDATA)>
<!ATTLIST dc:rights xml:lang CDATA #IMPLIED>
<!ATTLIST dc:rights rdf:resource CDATA #IMPLIED>
```

**Figure 2. The XML DTD of the DCMES XML binding**

[2] http://xml.apache.org/xindice
[3] http://www.softwareag.com/tamino
[4] http://www.ipedo.com
[5] see also section 5.

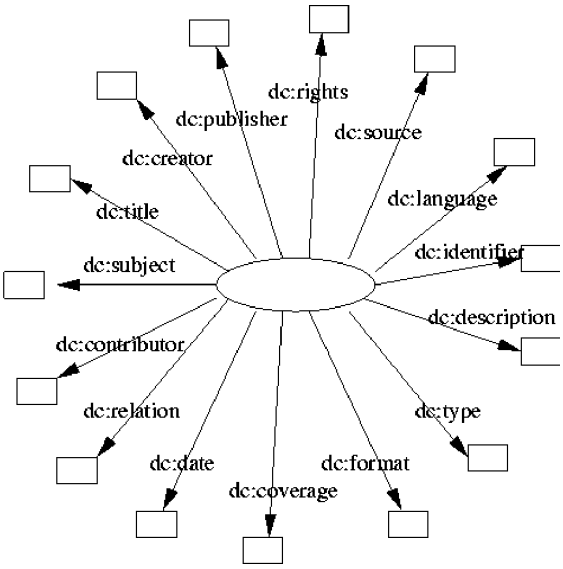



**Figure 3. The hedgehog model of the DCMES XML binding**

| | | |
|---|---|---|
| *R1: //*[@rdf:about] as u1* | ⇒ | Subject |
| *R2: u1/* as u2* | ⇒ | Predicate |
| *R3: u2[@rdf:resource] or u2[text()]* | ⇒ | Object |

**Figure 4. The rules used to map the DCMES XML binding data model onto ECDM's binary relational data model**

fixed resource, we can see that there exists an obvious mapping approach from the DCMES XML binding data schema to ECDM's binary relational data model. Moreover, since the DCMES XML binding only uses limited sets of RDF constructs (e.g., rdf:Bag, rdf:Seq, and rdf:Alt are excluded), the mapping becomes more straightforward. In Figure 4 we list three rules used to map the DCMES XML binding data model onto ECDM's binary relational data model. The XML data model is expressed here through XPath location paths using XPath's abbreviated syntax.

Note that in the DCMES XML binding data schema, the value of an element can be either plain text or another resource with a URI. This definition complies with the RDF data model and can be also appropriately expressed using XPath. In fact, based on the above mapping rules, the wrapper program can easily generate the ECDM-based common data view of the native XML repositories containing the DCMES XML binding metadata.

## 4. Translating Datalog into XPath

According to the usage scenario of the Edutella provider integration architecture, a common behaviour of the provider wrapper programs is to translate RDF-QEL queries into ECDM's internal query language Datalog. In addition, each wrapper program also has a specific behaviour: translating Datalog into the local query languages of the underlying repositories. Since the common behaviour of the wrapper programs has already been discussed in our previous publication [16], here we mainly concentrate on the specific behaviour of the wrapper programs of the native XML repositories, namely, translating ECDM's internal non-recursive Datalog queries, which correspond to the RDF-QEL sets from RDF-QEL1 to RDF-QEL3, into XPath.

Datalog is a non-procedural query language based on Horn clauses without function symbols. The basic construct of Datalog is the Atom, which describes ground assertion and can be represented in a simplified form corresponding to the binary relational data model as: *P(arg1, arg2)*, where *P* is Predicate that might be a relation name or arithmetic predicates (e.g., "<", ">", etc.), and *arg1*, *arg2* are Arguments that might be variables or constants. In Datalog, an Atom can be negated and represented as: *NOT P(arg1, arg2)*.

A Datalog program can be expressed as a set of Datalog rules. Each Datalog rule has a general representation as *head :- atom1, atom2,..., atomn*, where *head* is a single positive Atom, and *atom1* to *atomn* are a set of Atoms conjunctively called the body of the Datalog rule. Note that a Datalog rule may include negated Atoms in its body, but generally in some restricted forms [10]. Additionally, the disjunction in Datalog is expressed as a set of rules with the identical head. As an example, in Figure 5 we show a

```
H(X) :- P1(X,U), NOT P2(X, V)
H(X) :- P3 (X,W)

H is head; P1, P2, P3 are predicates; X is variable;
U, V, W are constants.
```

**Figure 5. A Datalog example query covering conjunctive query, disjunctive query, and query negation**

```
//*[@rdf:about and
    (P1 [@rdf:resource=U] or P1 [text()=U])  and
   not (P2 [@rdf:resource=V] or P2 [text()=V])
  ]
  |
//*[@rdf:about and
    (P3[@rdf:resource=W] or P3 [text()=W])
  ]
```

**Figure 6. A translated XPath query covering conjunctive query, disjunctive query, and query negation**

Datalog example query against a binary relational data model, covering conjunctive query, disjunctive query, and query negation. It corresponds to a typical RDF-QEL3 query defined in Edutella.

In Figure 6 we show the XPath query that is translated from the Datalog query illustrated in figure 5. As we have mentioned, the XPath expressions are based on the DCMES XML binding schema illustrated in Figure 2.

XPath can be seen as a general purpose query notation for addressing and filtering the elements and text of XML documents. A notation indicates the hierarchical relationship between the nodes and is used by a pattern to describe the types of nodes to match. All XPath queries occur within a particular context, which is the single node against which the pattern matching operates. The collections of all elements selected from the current context by XPath queries preserve document order, hierarchy, and identity, to the extent that these are defined. In addition, constraints and branching can be applied to any collection by adding a filter clause to the collection. The filter in XPath is analogous to the SQL WHERE clause, expressed in the form of *[filter pattern]*. The filter pattern evaluates to a Boolean value and is tested for each element in the collection. Any elements in the collection failing the filter pattern test are omitted from the result collection.

In general, each Datalog rule is mapped onto an XPath pattern, based on which a set of elements are selected under a certain context. The conjunctive queries, represented in Datalog by a number of Datalog Atoms contained in a single rule, are translated into sets of filter patterns that are combined together using the XPath Boolean operator *"and"* and are applied to the collection selected by the XPath pattern. The negation of a Datalog Atom can be represented using the XPath Boolean operator *"not"*.

The disjunctive queries, represented in Datalog by a number of Datalog rules with the identical head, are expressed in XPath by a number of patterns combined together using the XPath union operator "|". Multiple union operators can union together sets of collections selected by multiple XPath patterns, also being able to exclude duplicates. In the XPath query listed in Figure 6 we have also used several XPath operators for grouping operation, filtering operation, Boolean operation, and path operation. These operators are used according to certain precedence orders. In Table 1 we list these operators according to their precedence orders, from the highest to the lowest.

**Table 1. XPath operators and their precedence orders**

| | | |
|---|---|---|
| 1 | ( ) | Grouping |
| 2 | [ ] | Filter |
| 3 | / // | Path operations |
| 4 | \| | Union |
| 5 | not ( ) | Boolean not |
| 6 | and | Boolean and |
| 7 | or | Boolean or |

Note that against a binary relational data model, the example query listed in Figure 5 can be seen as a query for Subjects. In fact, in Datalog it is easy to express the queries for Predicates and Objects. Referring to the XPath expressions listed in figure 6, we can easily translate these Datalog queries into the corresponding XPath queries.

## 5. Adapting local query results into the Edutella common result exchange format

In Edutella, we have defined a RDF-based common result exchange format that represents query results as a set of tuples of variables with their bindings [16]. Whereas the RDF-based repositories can naturally adapt the local query results into Edutella's RDF-based common result exchange format with the support of some powerful local query languages, the native XML repositories can only return XML fragments selected by the XPath expressions. Regarding the DCMES XML binding schema based native XML repositories, the XPath queries can only return sets of whole XML metadata profiles that describe learning resources, since any XPath query must take the entire XML metadata profile as a whole in order to get a virtual binary relational data model against which the XPath query can be operated. Although most of native XML database implementations also provide means for further identifying the underlying elements/attributes of any XML fragments, we decided to use the whole XML metadata profile as the direct output and leave the further processing work on query results to a RDF parser, the Jena RDF Toolkit[6]. An important reason for this choice lies in the fact that the DCMES XML binding metadata profiles themselves are in valid RDF syntax and can be easily handled by RDF parsers. Through using the Jena RDF Toolkit, the query results generated by the native XML repositories can be easily transformed into the RDF model and then naturally adapted into the Edutella common result exchange format. However, it should be noted that in comparison to the query results returned from the RDF-based repositories, the query results from the native XML repositories are a bit redundant.

## 6. Integrating the LOM/IMS and SCORM XML binding schema based native XML repositories into Edutella

LOM is a learning resource metadata specification proposed by IEEE LTSC (Learning Technology

[6] http://www.hpl.hp.com/semweb/jena-top.html

Standards Committee), which specifies a conceptual data schema that defines the structure of a metadata instance for a learning object [11]. The LOM data schema is actually the basis of some other popular learning resource metadata specifications. For example, the IMS Learning Resource Metadata Specification directly employs the LOM data model and further provides an XML binding for it[7], the SCORM metadata specification extends LOM a little bit and provides a 100% downwards compatibility with it. In the following, our discussion will be based on the native XML repositories containing the LOM/IMS XML binding metadata. The SCORM XML binding schema based native XML repositories can use the same approach to be integrated into the Edutella network.

In comparison to the DCMES XML binding, the LOM/IMS XML binding data schema is much more complex, consisting of nine categories, over 50 metadata entries, and possibly recursive hierarchies (e.g., in the category "Classification"). In general, for such a complex XML schema, it is difficult to generate the ECDM-based common data view using XPath and further apply the same integration approach that is applicable to the DCMES based native XML repositories, as described in section 3, 4, and 5. At present some native XML database implementations begin to support a more powerful query language W3C XQuery [3], which provides a new possibility to generate the ECDM-based common data view of the LOM/IMS based native XML repositories and further apply the same integration approach. However, we argue that the XQuery-enabled new integration approach is more expensive than directly constructing the RDF-based metadata repositories using the LOM/IMS RDF binding [12] and further integrating these repositories into the Edutella network. In fact, for some complex learning resource metadata sets such as LOM/IMS and SCORM, using RDF is a more efficient and more extendible way for representing learning resources. Obviously, such types of repositories can be also more easily and naturally integrated into Edutella.

In order to address the immediate need of integrating the LOM/IMS based native XML repositories into the Edutella network, we employ the approach that relies on the DCMES XML binding as a lingua franca and scale-down maps the LOM/IMS XML binding into the DCMES XML binding through using W3C XSLT (XML Stylesheet Language Transformations) [5]. After the transformation, the integration approach for the DCMES-based native XML repositories can be directly applied to the LOM/IMS based native XML repositories.

As one can imagine, such a transformation from LOM/IMS to DCMES unavoidably loses some information of the original LOM/IMS metadata set. However, we argue that most of lost metadata information are useful only for detailed description of learning resources rather than for the simple discovery of these resources. Thus our integration approach for the LOM/IMS based native XML repositories can still ensure the essential discoverability of the learning resource metadata contained in these repositories. Moreover, the validity of this integration approach is also guaranteed by the common efforts from IEEE LTSC and DCMI (especially the Dublin Core Education Working Group[8]), which have been continuously focused on providing enough interoperability between LOM/IMS and DCMES, as outlined in the MoU[9] (Memorandum of Understanding) between IEEE LTSC and DCMI.

In Table 2 we list the 15 rules used to map LOM/IMS to DCMES [11]. Based on these rules, the transformation from the LOM/IMS XML binding to

**Table 2. The rules used to map LOM/IMS to DCMES**

| LOM/IMS | DCMES |
|---|---|
| 1.1.2:General.Identifier.Entry | DC.Identifier |
| 1.2:General.Title | DC.Title |
| 1.3:General.Language | DC.Language |
| 1.4:General.Description | DC.Description |
| 1.5:General.Keyword or 9:Classification with 9.1: Classification.Purpose equals "Discipline" or "Idea". | DC.Subject |
| 1.6:General.Coverage | DC.Coverage |
| 5.2:Educational.Learning ResourceType | DC.Type |
| 2.3.3:LifeCycle.Contribute.Date when 2.3.1: LifeCycle.Contribute. Role has a value of "Publisher". | DC.Date |
| 2.3.2:LifeCycle.Contribute. Entity when 2.3.1: LifeCycle. Contribute.Role has a value of "Author". | DC.Creator |
| 2.3.2:LifeCycle.Contribute.Entity with the type of contribution specified in 2.3.1: LifeCycle. Contribute.Role. | DC.Other-Contributor |
| 2.3.2:LifeCycle.Contribute. Entity when 2.3.1: LifeCycle. Contribute.Role has a value of "Publisher". | DC.Publisher |
| 4.1:Technical.Format | DC.Format |
| 6.3:Rights.Description | DC.Rights |
| 7.2.2:Relation.Resource.Description | DC.Relation |
| 7.2:Relation.Resource when the value of 7.1:Relation.Kind is "IsBasedOn". | DC.Source |

[7] until now IEEE LTSC itself has not yet provided the XML binding for LOM.
[8] http://dublincore.org/groups/education/
[9] http://dublincore.org/documents/2000/12/06/dcmi-ieee-mou/

the DCMES XML binding can be easily accomplished by an XSLT program.

In the native XML repositories, all XML metadata profiles are stored in the separate XML collections according to certain XML schemas. Utilizing an XSLT program, we can easily create a specific collection to store the transformed LOM/IMS metadata profiles, just like creating a database view in RDBs. Moreover, since each XML metadata profile stored in the native XML repositories possesses a unique key to identify itself, we can also retrieve the original metadata profile and get all metadata information.

## 7. Conclusions

Due to the considerable incomparability between the XML data model and the RDF data model, a generic approach for integrating schema-agnostic native XML repositories into the RDF-based Edutella network was deemed to be unrealistic for our application. This is also attributable to the fact that XPath, the local query language of the native XML repositories, is less powerful and thus incapable of manipulating some complex XML data models to generate their ECDM-based common data view. Moreover, XPath is also incomparable to ECDM's internal query language Datalog and thus incapable of supporting full relational algebra queries. At present, some native XML database implementations begin to support a more powerful query language W3C XQuery, which provides a new possibility to manipulate the native XML repositories and is also more comparable to Datalog (besides providing additional features for handling and creating hierarchical data structures). However, we argue that for schema-agnostic native XML repositories, integrating them into Edutella through using XQuery is more expensive than the integration approach of constructing the RDF-based repositories and then directly integrating them into Edutella. As a matter of fact, for some complex metadata sets such as LOM/IMS and SCORM, using RDF and some high-performance back-end systems is a more efficient and more extendable way for building learning resource metadata repositories. Therefore, although we have found a feasible approach for integrating schema-specific native XML repositories into the Edutella network, which has satisfactorily addressed the current usage and immediate integration need of Edutella by covering most of popular learning resource metadata sets such as DCMES, LOM/IMS, and SCORM, we encourage the application of more RDF-based learning resource metadata repositories in the Edutella network, given the inherent advantages of RDF in distributed P2P settings, such as the easy composability of schemas, as well as the extendability and modularity of distributed RDF metadata.

## References

[1] ADL Technical Team, SCORM Specification V1.2, http://www.adlnet.org/index.cfm?fuseaction=scormat

[2] Beckett, D., E. Miller, and D. Brickley, Expressing Simple Dublin Core in RDF/XML, http://dublincore.org/documents/2001/11/28/dcmes-xml/

[3] Boag, S., D. Chamberlin, M. F. Fernandez, D. Florescu, J. Robie, J. Siméon, and M. Stefanescu, XQuery 1.0: An XML Query Language, http://www.w3.org/TR/2002/WD-xquery-20020430/

[4] Brantner, S., T. Enzi, S. Guth, G. Neumann, and B. Simon, UNIVERSAL - Design and Implementation of a Highly Flexible E-Market Place of Learning Resources, in Proc. of the 1st IEEE International Conference on Advanced Learning Technologies (IEEE ICALT 2001), Madison, WI, USA, Aug. 2001.

[5] Clark, J., XSL Transformations (XSLT) Version 1.0, http://www.w3.org/TR/xslt

[6] Clark, J., and S. DeRose, XML Path Language (XPath), http://www.w3.org/TR/xpath

[7] DCMI, Dublin Core Metadata Element Set, Version 1.1, http://dublincore.org/documents/1999/07/02/dces

[8] Dhraief, H., W. Nejdl, B. Wolf, and M. Wolpers, Open Learning Repositories and Metadata Modelling, in Proc. of the International Semantic Web Working Symposium, Stanford, CA, USA, Aug. 2001.

[9] Duval, E., E. Forte, K. Cardinaels, B. Verhoeven, R. Van Durm, K. Hendrikx, M. Wentland-Forte, N. Ebel, M. Macowicz, K. Warkentyne, and F. Haenni, The ARIADNE Knowledge Pool System, Communications of the ACM, 44 (5), 2001.

[10] Garcia-Molina, H., J. D. Ullman, and J. Widom, Database Systems: The Complete Book, Prentice Hall, USA, 2001.

[11] IEEE LTSC, IEEE LOM working draft 6.4, http://ltsc.ieee.org/wg12/index.html

[12] IMS Global Learning Consortium, IMS Learning Resource Metadata Specification V1.2.1, http://www.imsproject.org/metadata/index.html

[13] Kokkelink, S., and R. Schwänzl, Expressing Qualified Dublin Core in RDF/XML, http://dublincore.org/doc uments/2002/04/14/dcq-rdf-xml

[14] Lassila, O., and R. R. Swick, Resource Description Framework (RDF) Model and Syntax Specifica-

tion, http://www.w3.org/TR/REC-rdf-syntax

[15] Liu, X., K. Maly, M. Zubair, and M. L. Nelson, Arc - An OAI Service Provider for Cross Archiving Searching, in Proc. of the ACM/IEEE Joint Conference on Digital Libraries, Roanoke, VA, USA, June 2001.

[16] Nejdl, W., B. Wolf, C. Qu, S. Decker, M. Stintek, A. Naeve, M. Nilsson, M. Palmer, and T. Risch, Edutella: A P2P Networking Infrastructure Based on RDF, in Proc. of the 11th International World Wide Web Conference (WWW 2002), Hawaii, USA, May 2002.

[17] Qu, C., and W. Nejdl, Towards Interoperability and Reusability of Learning Resources: a SCORM-conformant Courseware for Computer Science Education, in Proc. of the 2nd IEEE International Conference on Advanced Learning Technologies (IEEE ICALT 2002), Kazan, Tatarstan, Russia, Sept. 2002.

[18] Papakonstantinou, Y., H. Garcia-Molina, and J. Widom, Object Exchange Across Heterogeneous Information Sources, in Proc. of the 11th International Conference on Data Engineering, Taipei, Taiwan, March 1995.

[19] Teachware on Demand,
http://www.teachwareondemand.de

[20] Thompson, H. S., D. Beech, M. Maloney, and N. Mendelsohn, XML Schema Part 1: Structures, http://www.w3.org/TR/xmlschema-1

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 91-96*


# Building Digital Books with Dublin Core and IMS Content Packaging

Michael Magee
Netera Alliance
magee@ucalgary.ca

D'Arcy Norman, Julian Wood, Rob Purdy, Graeme Irwin
University of Calgary
dnorman@ucalgary.ca, woodj@ucalgary.ca, rpurdy@ucalgary.ca, irwing@cpsc.ucalgary.ca

## Abstract

*The University of Calgary Learning Commons has been developing solutions for the digitization and dissemination of educational digital assets for two years. The most recent work focused on creating digital books that can be accessed online and assembled from digital components. Several theoretical and technical issues were examined and resolved. The U of C worked with existing partners to modify their educational object repository software solution to meet these needs. The software was developed to deal with the workflow of assembling the numerous digital components of a book into a cohesive whole and an online browser was built to view the constructed digital books. The digital books were created in an XML-based IMS container package of Dublin Core metadata and manifests of all the components that were used to create the online digital books.*
**Keywords:** *Dublin Core, Digital Books, IMS, Metadata, XML*

## Introduction

The efficient management of information is a driving force behind modern society. Canada is not immune from this trend. The Federal government looks upon the Internet as an opportunity to "fulfill its responsibilities in the generation and dissemination of information in a more effective and timely manner" [4]. This has led to a number of recommendations to enable the vision of creating an accessible body of digital content for Canadians and the rest of the world.

A large part of this mandate has been the digitization or re-purposing of existing content. The University of Calgary has been working with the CAREO (Campus Alberta Repository of Educational Objects) and the BELLE (Broadband Enabled Lifelong Learning Environment) projects to take existing educational content and place it online. As the body of existing content undergoes review it becomes obvious that a framework needs to be in place to deal with decisions of what content will be chosen for digitization and how it needs to be re-organized to work in an online environment. The spectrum of media that is a candidate for movement into the digital realm is considerable. These include text, video, film, photographs and a host of new and emerging multimedia formats. Books represent one of the oldest and one of the most challenging of these formats.

The intellectual cull of the vast herd of literature that currently exists is not an easy decision. As books are deigned worthy of preservation the ones that do not meet the criterion are lost to future generations forever. Other decisions may not affect the preservation of a resource but it will affect its accessibility to the general public, as more and more dependence is placed on online resources alone. The decisions are therefore not to be taken lightly as they may represent an intellectual and sociological bias that will affect our worldview.

Our project was not initially focused on the decisions about what books would be chosen for digitization. We had to examine how to create digital books so that they would be easy to find, accessible and of great utility. We saw the opportunity to explore a new approach to the creation of digital books. The IMS content package is an abstract container designed to describe a large, complex hierarchical data structure as well as its component media. It was designed to allow the movement of complex objects between systems and give them the ability to communicate with other software. The solution we chose is only one of many options available in the world of information technology and will need to be evaluated to determine its appropriateness as a new tool. It represents

a solution that will combine the approaches currently being used in the world of educational technology and library science. The solution has a number of benefits that will create a digital book that is not only readable but has considerably more utility that just the book on its own.

This is an important consideration as there is a movement towards efficiency in the library world that will attempt to increase effectiveness by ending the physical book and replacing those anachronisms with pure data [5]. There need to be as many technological options as possible for the book to be properly evaluated. There is no doubt that many books will lack utility in some areas but will be extremely useful in others. The largest number of options available to those archiving and sharing those volumes will provide the most justice to the choices being made about which volumes will chosen.

## Previous Research

There was a considerable amount of research that had occurred to examine the issues surrounding the search, retrieval and organization of educational digital assets online. This work became the basis for development of the mechanisms to deal with more complex organization of objects.

## CAREO

The CAREO project is involved in the research and development of both a provincial and a national educational object repository [2]. Educational objects can be defined in a variety of ways but there are a few common elements. They are fundamentally small, digital instructional components that can be reused a number of times in different learning contexts and delivered over the Internet [10].

The system was created to address the problems of the explosion of online, digital educational content and the increasing difficulty in locating and utilizing that content.

As a result of this research CAREO developing a networked repository system that displayed XML document records based on IMS metadata, an educational metadata set. Although the focus was on educational applications the architecture was kept as flexible as possible so that any metadata standard based on an XML-schema could be stored.

The CAREO application software is designed to allow the search, retrieval and display of IMS metadata records in a web browser. These records are linked to educational objects located online. The CAREO repository is designed to be a modular component of a larger system (Figure 1). It has a built in communication layer based on XML-RPC that allows other repositories and tools to search and utilize the features of the software. In its current implementation, CAREO is integrated with the ALOHA (Advanced Learning Object Hub Application) metadata server which provides additional functionality in the role of a middleware layer between the CAREO application and the user's browser application.

For the user, the CAREO application acts as an educational portal or website, providing a central point of reference for educators and students when looking for information and resources to support teaching and learning. By providing a set of tools to enable such activities as resource discovery (searching and browsing), publication, aggregation, and sharing, CAREO is able to provide meaningful and immediate access to online materials.

By implementing the IMS Content Packaging specification, CAREO has been able to extend its suite of tools to enable its users to create compound aggregations of learning resources. These compound aggregations may range from simple collections of images into a single package, to electronic representation of physical books, to highly structured online courses.

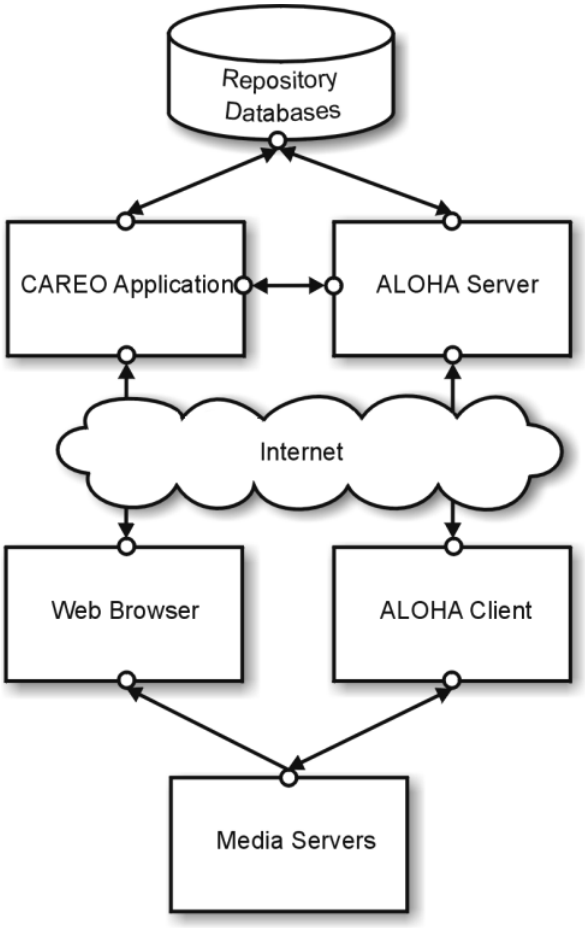



**Figure 1. CAREO/ALOHA Architecture**

## ALOHA

The ALOHA java client was developed as the result of research early in the CAREO project. It indicated that the time and effort required to create metadata records for the individual educational objects was much too long for most of the projects and academics involved. A tool to streamline the upload of metadata and the media itself was therefore required to improve the workflow of objects into the system. The ALOHA tool is a java client that was designed to allowing indexing using any metadata standard. The decision to use Java was based on the power and flexibility Java has demonstrated in interacting with the World Wide Web [6]. The tool ingests a valid XML schema, creates a data entry interface based upon the schema and allows both amateur users and professional indexers all the simplicity and sophistication they require. It is easy to create, share and customize indexing templates and forms (Figure 2).

It also integrates drag and drop functionality that can automatically extract metadata from over 200 files types. It makes marking up IMS, or other forms of metadata, much easier. Administrative tools managing workflow issues with multiple indexers including the librarian, the educator, and the media developer are available. This supports the idea of modularity where different users can index objects in context specific ways and share their metadata with other users and metadata schemas [3]. Once the indexing is complete the media and metadata can be uploaded simultaneously with the touch of a button to an appropriate media-server, handling the job of an FTP program.


**Figure 2. ALOHA Interface**

## The Problem

A search and retrieval system had been successfully implemented for educational objects but as more complex objects were examined it became apparent that a greater degree of sophistication would be necessary to deal with large constructs. Books were one of the most obvious assets the system would be unable to handle. The current system could handle a simple description of the book as a single entity. Unfortunately books, like many types of complex media, are composed of a large number of organizational structures such as sections and chapters that combine to make up their whole structure.

It was necessary to retain the search and retrieval features of the metadata but there was also a need to describe the actual structure of the object so that it could be assembled for online viewing in a meaningful way.

## The Solution

The first physical asset chosen to test structural metadata was a book. Simplistically, books can be thought of as hierarchical arrangements of content. Words are aggregated into sentences, sentences into pages, pages into chapters, and chapters into books. In constructing digital books it was convenient to follow this pattern. The CAREO project wanted to continue to develop and support IMS metadata technology and therefore looked at IMS Content Packages as a solution. This involved treating the pages, chapters and books of the digital books as IMS content packages. There have been other solutions to the problems of digital books. The Library of Congress has taken a similar approach using METS (Metadata Encoding and Transmission Standard) [7]. The Open eBook standard has also been created to describe digital books in a standardized way [8]. In all cases structural metadata was created to describe the separate digital files that could make up a book

## IMS Content Packaging

The IMS Content Packaging Specification 1.1.2 was designed to assist in the creation of complex educational content [1]. Basically, a content package allows numerous assets to be brought together, organized, and described. The educational objects within the package can have several organizational structures so that one content package can place them in a number of different contexts, educational or otherwise.

At the top level of the IMS content package is the Manifest (Figure 3). A Manifest consists of Resources, Organizations, Metadata, and optional sub-Manifests. The Resources are the actual digital files and/or links that makeup the package content.

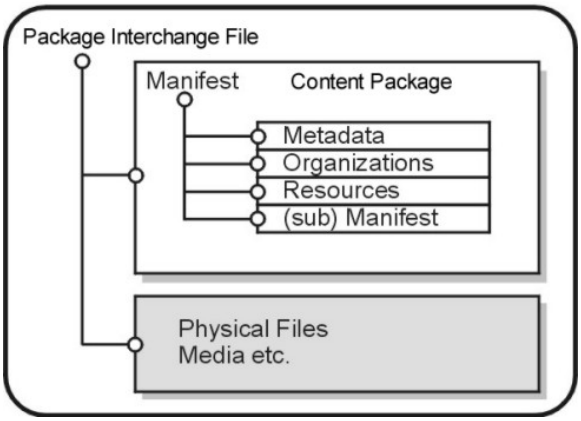



**Figure 3. IMS Package Interchange File**

The Organizations section is a hierarchical arrangement of the Resources. This is where the content is given order and context. The Metadata section provides an area to add descriptive metadata for the content package. Additionally, metadata can be optionally attached to nearly any part of the Manifest. It is important to note that this metadata can follow any metadata standard. Although IMS recommends that its own IMS metadata specification be used it is not required

An example of how a content package might be used to construct a book follows. First, a content package representing a chapter could be created. The pages that belong to that chapter are then added as Resources. Each page is put in proper order using the Organizations section and metadata could be added to each page to further describe it. This process can be repeated for each chapter in the book. Then, a content package representing the book could be created and each chapter added, ordered, and annotated.

The treatment of each part of the book as a separate package allows each section to be independently manipulated, searched, and created. The specification would also allow the XML file describing the manifest and all the physical files used in the package to be bundled together into a Package Interchange File that could be compressed into a single file and moved between systems. As the standard was published a number of vendors were creating software that could both build, package and exchange these Package Interchange Files. In the educational world this offered the opportunity to share not only atomic educational objects but also large aggregations that could be formed into lessons, courses and entire programs.

## Test Case: Canadian Local Histories

The "Our Roots/Nos Racines" project is a project that was initially undertaken by the University of Calgary and Laval University. The project is doing an inventory and assessment of all media associated with Canadian cultural heritage. The goal is to get as much of that material into an online venue that is accessible to all Canadians [9]. Initially the library projects wanted to take the thousands of pages of local histories it had digitized and place them online as complete books for online access by genealogists and researchers. The first phase of the project created digital, online versions of local histories by scanning the books into graphic files and uncorrected OCR files. The combination of the two allowed for rough text searching and display of the actual page in a web browser.

Under the direction of Tim Au Yeung at the University of Calgary the initial proof of concept was successful but it became apparent that the system was going to need to scaled up to accommodate a number of other repositories, many more types of digital assets, larger volumes and data and increase in users. The Our Roots project consulted the CAREO project as it was involved in research and development in this area.

There were a number of similarities between the needs of the two projects. Both of them needed to research and develop ways of organizing and structuring large volumes of online content. Where they differed was in the type of metadata being used to describe the content packages. The library project was using Dublin Core metadata and it's own proprietary extensions but both systems had metadata as a focus if a search and retrieval system as a common element. The use of Dublin Core in the IMS Content Package did not represent a difficulty as the container package was designed to be generic and flexible enough that it could contain many kinds of descriptive metadata.

The limitation of the IMS Content Package in describing books was the generic nature that made it so useful in the first place. Other digital book standards were designed explicitly to describe books while the IMS standard could describe any combination of digital assets. This required that the structure had to be carefully defined while the books were being assembled. It was critical that the packages identify themselves as pages, chapters and books as there was no pre-defined slot for those elements.

The IMS Content Packages were assembled in the ALOHA software from scanned components of the book. These included the text from the OCR and several sizes and formats of digital images of the page itself. ALOHA would treat them like digital assets and allow the users to organize them into pages, chapters and books. Once organized, the files were moved online. The digital files representing the pages of the books were moved to the media servers and the Dublin Core metadata representing the description of the various components of the book was moved to the metadata server. The CAREO software was used to display the digital books online.

CAREO used a browsing structure that allowed the books to be browsed, searched and read online (Figure 4). The search and retrieval aspect of the Dublin Core metadata used to describe the various components of the book would allow searching down to the level of page in the book.

## Conclusion

IMS Content Packing presents one of a number of XML-based container package standards for digital books. The advantages of the IMS standard come from the generic nature of the content package. A book can be one of several media types all within the same package. This allows a book, chapter or page to be part of a larger, more complex multimedia presentation. This opens books up to a larger realm of opportunity than just the library.

The work of education is demanding the creation of many new Learning Management Systems based on IMS standards. These systems will be able to import and export IMS Package Interchange Files and present the IMS Content Packages to the students and teachers. When books are described using this standard these systems will be able to ingest a page, a chapter or a whole book as part of a course or a lesson. The packages could expand to include lessons and tests specific to a book and target audience.

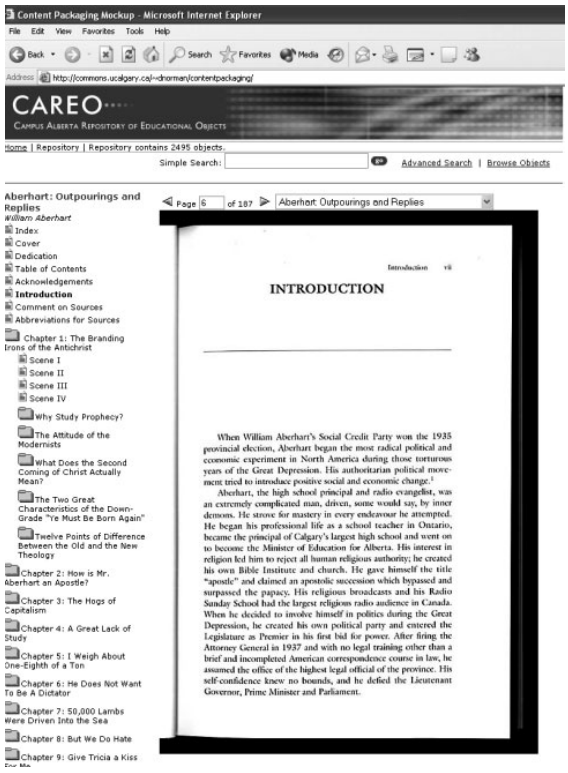



**Figure 4. CAREO Book Browsing Interface**

The packages will also have the ability to communicate with a LMS through a standardized API. This will allow instructors to track progress through the book and the test and score results of a student working through the book online. As more and more vendors create software and tools that can work with the IMS standard content that is described in this way will gain access to a greater level of utility and interoperability

The same ability that allows content packages to be exported as interchange files would make it possible to easily move digital books between libraries. This movement could be just the organization of the book linked to its online components or a complete package that included its organizational structure and all its digital assets. The generic nature of the metadata used to describe contents of the package allows the use of Dublin Core as well as other metadata standards to describe the components of the package. It would be possible to add additional metadata sets as well so that a book that was using Dublin Core for search and retrieval metadata could also add a section of IMS metadata that would describe its educational context.

The storage of the digital book IMS Content Package Information in XML provides opportunities to move the data between standard digital book formats. There is also a degree of similarity in the structure of other digital book packaging standards that could eventually allow a degree of interoperability between books stored in the various formats. The potential to move books between these formats and allow them a large venue of exposure in many different contexts is a definite avenue for future research in this area.

## References

[1] Anderson, Thor & McKell, Mark
*IMS Content Packaging Best Practice Guide: Version 1.1.2 Final Specification*,
http://www.imsproject.org/content/packaging/cpv1p1p2/imscp_bestv1p1p2.html, 2001.

[2] CAREO Project, CAREO project website, http://careo.netera.ca, 2002.

[3] Duval, E., Hodgins, W., Sutton, S., and Weibel, S., Metadata Principles and Practicalities. *D-Lib Magazine*, Vol. 8, No. 4,
http://www.dlib.org/dlib/april02/weibel/04weibel.htm, 2002.

[4] Federal Task Force on Digitization, *Towards a Learning Nation: the Digital Contribution: Recommendations Proposed by Federal Task Force on Digitization*, Canadian Government Publication, Ottawa, 1997.

[5] Hannah, S.A. and Harris, M.H, *Inventing the Future: Information Services for a New Millennium*, Ablex Publishing Corporation, Stamford, Connecticut, 1999.

[6] Jones, P, Java and Libraries: Digital and Otherwise. *D-Lib Magazine*, http://www.dlib.org/dlib/march97/03jones.html, March 1997.

[7] METS Metadata Encoding and Transmission Standard, *Metadata Encoding and Transmission Standard Official Web Site*, http://www.loc.gov/standards/mets/, 2002.

[8] Open eBook Forum, http://www.openebook.org/, 2002.

[9] Our Roots, Nos Racines, http://ahdptest.lib.ucalgary.ca/CDLHS/, 2002.

[10] Wiley, D.A., Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy, In D. A.Wiley (Ed.), *The Instructional Use of Learning Objects: Online Version*, http://reusability.org/read/chapters/wiley.doc, 2000.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 97-103*


# The Virtual Image in Streaming Video Indexing

Piera Palma, Luca Petraglia, Gennaro Petraglia
Dipartimento di Matematica e Informatica - University of Salerno (Italy)
I-84081 Baronissi, Salerno, Italy
petragli@unisa.it
tel +39 089 965248 fax +39 089 965438

## Abstract

*Multimedia technology has been applied to many types of applications and the great amount of multimedia data need to be indexed. Especially the usage of digital video data is very popular today.*

*In particular video browsing is a necessary activity in many kinds of knowledge. For effective and interactive exploration of large digital video archives there is a need to index the videos using their visual, audio and textual data. In this paper, we focus on the visual and textual content of video for indexing.*

*In the former approach we use the Virtual Image and in the latter one we use the Dublin Core Metadata, opportunely extended and multilayered for the video browsing and indexing.*

*Before to concentrate our attemption on the visual content we will explain main methods to video segmentation and annotation, in order to introduce the steps for video keyfeature extraction and video description generation.*

**Keywords:** *Video and Image Indexing, Video Browsing, Keyframe, DC Metadata , Virtual Image.*

## 1. Introduction

Digital video is becoming the rising tide of multimedia. The amount of video data is growing dramatically. Thus indexing and cataloguing of digital videos are more and more important for retrieval. The best way for indexing video data is content based. In the past, we usually described and annotated video content manually. However this traditional solution is not suitable for the enormous amount of video data. We must find a mechanism that can provide an efficient and flexible solution to illustrate video content. In order to analyse video content we must to segment its content in units. It is possible to do this at two levels:

- *Structural level*, and then we divide videos into frames, shots, clips, episodes or scenes;
- *Content level*, according to cinematographic properties, motion of the camera, audio properties, motion of a character/object, scenes and stories within a video, etc.

This paper is organized as follows. In section 2 we describe the two levels of video analysis mentioned above. In section 3 we introduce the criteria of choice for metadata to video indexing and how we apply these metadata to video segments used in our processes of video indexing. In section 4 we describe the Virtual Image and in section 5 we say *why* we use it to video indexing and *how* this content based method can manage also the metadata. In section 6 we make our conclusion on the work.

## 2. Video Segmentation and Video Extraction/Annotation

Indexing on video content is possible from two points of view: *temporal segmentation* and *content analysis*. The first is the identification of meaningful video segments (as shots, scenes, and episodes); the second is the identification of attributes characterizing regions, objects, motions in a video segment. We briefly describe both below. We define segmentation the process of breaking down a video into its constituent basic elements, that is the shots, and their higher-level aggregates, such as episodes or scenes. There are traditional approaches to performing segmentation composed by the following steps: previewing the whole video, identifying the shots, episodes and scenes and then providing them and their boundaries of textual labels. Since this solution is very time-consuming there is a less expensive way, that is to use the *edit decision list* created by video producers during post-production, but there are few producers that use this method. The detection of shot boundaries is possible either on the raw video stream or on compressed data. There are two main methods to do this:

- Cuts detection, where the cut is defined as a clean transition between a shot and the following; it generally corresponds to a curt change in the brightness pattern of two consecutive images;
- Gradual transitions detection, where the change from one shot to another is detected through a number of frames which present some optical effect as fade-in and fade-out, wipes and mattes, etc.

Since a typical segmentation into shots of some types of video (like movies, news and documentaries) produces too many shots (e.g. 600-1500 in a movie) there is the need to build shot aggregates, useful not only for the evaluation of video content, but also for video access at semantic level; for example a sequence of short shots stresses fast action while a sequence of shots with motion, alternated with static shots, stresses dynamics. The shot can be an effective method to segment some formats of video, where it is a useful basis to create new episodes (e.g. in news video), but it is very laborious for video formats where the complete fruition process prevails (as in shot aggregates or episodes).

An important concept for the detection of shot aggregates is the *keyframe*, that is a particular frame from the video stream that represents its content or, more usually, a part of it. Higher level aggregates in a movie can be detected by analysing the similarity between keyframes or repetition of shot keyframes. An example of use of keyframe is in [13], where in order to create an automatic video content description, video is firstly segmented in scenes, that compose the *story unit*; keyframes are extracted from them and then *key features* are produced. Finally *descriptors* are generated. We summarize this process in Fig. 1.

Once a video stream is segmented into its constituent elements, it is necessary that content indexes are set. We create indexes on objects and motions, either on the meaning conveyed by visual primitives.

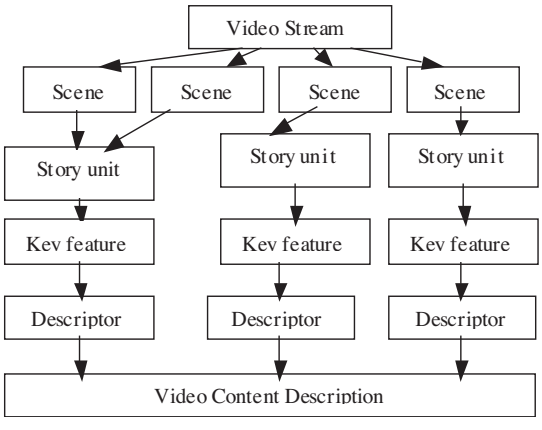



**Fig 1. The description generation**

Indexes on objects are usually extracted from the keyframe (as mentioned above); the keyfeatures (informations extracted from the keyframe) are used in comparison with primitives (or features) extracted from the query image. The indexes mentioned above are usually full text keywords, or a structured set of concepts, both obtained with human intervention. But it is also possible the use of algorithms in image analysis for automatic extraction of keyfeatures. Different types of video need different types of indexes on video content.

But we are interested in manual annotation and in particular in *visual iconic annotation*. It combines two distinct representations:

– A *semantic representation*, which is independent from temporal ordering of object actions;
– A *temporal representation* which etablishes specific relationship among objects through their combination and temporal ordering of their actions.

*Icon*s visually represent categories or situations that are in the video, used as visual primitives or compound descriptors. The annotation is usually based on visual languages. An approach particularly suited to describing object spatio-temporal relationships in a sequence is the *iconic annotation by example*, where visual examples are built; these visual examples represent the content of a video segment that will be parsed into a simbolic sentence, according to a special description language. This approach has been used by some authors for its expressiveness and because through it we can generate very detailed descriptions of dynamic content of a video stream. From these authors we mention Arndt and Chang [1] and Del Bimbo et al. [8]. Arndt and Chang have suggested symbolic description of motion trajectories for indexing video content through 2D Strings (to represent object spatial arrangement in individual frames) and set-theory (to describe changes due to motion).

Del Bimbo et al. presented the language Spatio Temporal Logic (STL) in order to represent in a symbolic way spatio-temporal relationship in shot sequences. The basic idea of STL is the *spatial assertion*, that captures the spatial arrangement of the objects in a scene. Groups of successive frames with equivalent spatial descriptions constitute the *states*, which in turn are combined through the Boolean connectives and the *temporal-until* operator. Finally the expression constructed with STL will be parsed in a visual sentence (this mechanim is particularly used in the querying phase).

## 3. Metadata in the video indexing process

Currently video indexing through the use of standard metadata caused a great interest from different research groups, among these the DCMI Moving Pictures Special Interest Group; on its proposal we will base ours. Firstly we need to define our criteria

to video segmentation (which we will derive from the analysis of some criteria seen in previous section). Afterwards we will propose for those levels (in which the video is segmented) the corresponding metadata, whose elements will be just derived from Dublin Core metadata element set. Our proposal on video segmentation is based on modification of scheme showed in fig.1, where two video segmentation levels surface: the first level is the *scene*; the second one is the *story unit*.

*Definition* A *story unit* is the aggregation of many scenes logically connected. It differs from the concept of sequence since scenes connected in a story unit can be also not contiguous, while in the sequence scenes are contiguous. Since such aggregation occurs only at logic level, story units are logical entities, which are constructed through the use of metadata.

The advantages of introduction of such entity are:

- It does not have to be phisically stored, but it need to be characterized in the system catalog of OODB. Consequently it will provide a greatest amount of informations without futher waste of storage;
- It is a logical aggregate of scenes and then can be characterized by a specific *Keyframe;*
- It can be defined through the use of metadata, and this approach can be extended also to key-frames and scenes;
- The indexing and querying processes use search engines based on metadata.

For entities that we chose the following levels of metadata are defined:

1. The first level is for metadata on the whole video (for it we adopt the classical approach using the whole set of Dublin Core metadata) and for the scene (for it we use a subset of the above-mentioned metadata), opportunely extended as specified in J. Hunter's proposal [24] (e.g. using description.keyframe, description.startTime, description.endTime, description.text);
2. The second level is for metadata on the story units, obtained using a small subset of extended Dublin Core metadata (we detail this level below);
3. A third level is for metadata on the keyframe (possibly based on clustering processes), that uses Virtual Image (described in detail in the next section);

In particular we will focus in the second level; for this one only the following metadata are necessary:

- ***Subject***: Since story units are created for cataloguing and fruition, this element functions as title and subject at the same time. In fact, while for the video a known title of the work usually exists, for the story units it does not exist; then in the story units we can to indicate the category (as action, dialogue, etc.)
- ***Description***: For this element we use the following extentions:
  – Description.Text
  – Description.Keyframe
- ***Type***: With it we indicate the type of resource between the possible ones for the video streaming (as video, scene, shot, frame, at which we add *story unit)*
- ***Relation***: This element is important since it implicitly allows to ***inherit*** from video the remaining Dublin Core metadata. In fact in the story units (and in the scenes that compose the story units) we use the descriptor ***Relation.IsPartOf***; it joins such entities to "father" video (the video from which we extract scenes and story units). Then we derive the remaining attributes from the "father" video. Moreover for the story units we propose the *Relation.HasPart* extention, in order to connect story unit with scenes whose it is composed

It is necessary to focus on the ***Description. Keyframe*** element, that represents story units and then scenes. It is just the beginning point of our *content&metadata based* cataloguing. Then we can modify the scheme of Fig. 2 as follows:

st the beginning point of our *content&metadata d* cataloguing. Then we can modify the scheme of as follows:

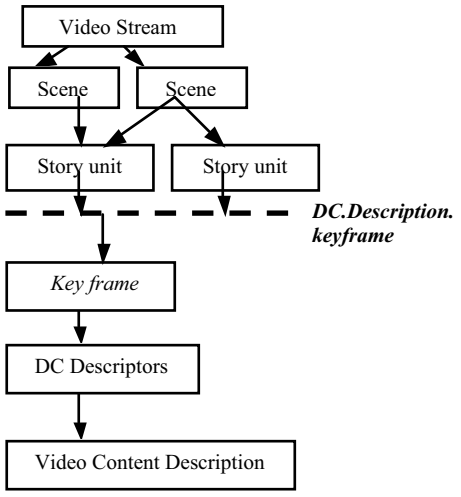



**Fig 2. The *metadata-based* video indexing**

## 4. The Virtual Image

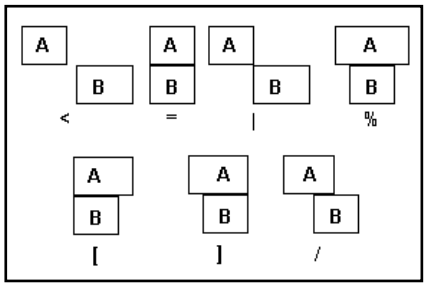
From original point of view the Virtual Image [17] describes its corresponding real image in terms of objects and spatial relationships and preserves the spatial knowledge embedded in the real image.

Formally it is defined as a pair (Ob, Rel) where :

- Ob = $\{ob_1, \ldots\ldots, ob_n\}$ is a set of objects
- Rel = $\{Rel_x, Rel_y\}$ is a couple of sets of binary spatial relationships on Ob, in particular $Rel_x$ (resp. $Rel_y$) contains disjoint subsets of Ob x Ob that

express spatial relationships "<" , "|","=", "[", "]", "/", "%", between object pairs of *im* (the real image) on *x* axis (resp. *y* axis)

For simplicity we use the notation $ob_i\ \gamma\ ob_j$ to indicate that the pair $(ob_i, ob_j)$ belongs to the relation γ, where $ob_i, ob_j \in Ob$ and $\gamma \in \{>, |, =, [, ], /, \%\}$. A triple like $ob_i\ \gamma\ ob_j$ is called an *atomic relation* in the following. We also say that atomic relation $ob_i\ \gamma\ ob_j$ *belongs to* $Rel_x$ (resp. $Rel_y$) if the spatial relation holding between $ob_i$ and $ob_j$ along the x - projection (resp. y – projection) is γ. We can regard both $Rel_x$ and $Rel_y$ simply as sets of atomic relations. In the figure below




we show possible spatial relations:

**Fig 3. Example of possible spatial relations betweeen A and B**

The Atomic Relation Extraction Method (AREM algorithm) derives Virtual Image from a given real image through the following steps:

**Step 1:** Let $Rel_x$ (resp. $Rel_y$) be empty set;

**Step 2:** Scan the image along the x-direction (resp. y-direction) to compute the values begin (A) and end (A) for every A∈ Ob;

**Step 3:** For each couple (A,B) ∈ Ob, add to $Rel_x$ (resp. $Rel_y$) the relation obtained by the following case-statement:

Case:

| | |
|---|---|
| end(A)<begin(B) | : A<B |
| end(B)<begin(A) | : B<A |
| begin(A)=begin(B) and end(A)=end(B) | : A=B |
| end(A)=begin(B) | : A\|B |
| end(B)=begin(A) | : B\|A |
| begin(A)<begin(B) and end(A)>end(B) | : A%B |
| begin(B)<begin(A) and end(B)>end(A) | : B%A |
| begin(A)=begin(B) and end(A)>end(B) | : A[B |
| begin(A)=begin(B) and end(A)<end(B) | : B[A |
| begin(A)<begin(B) and end(A)=end(B) | : A]B |
| begin(A)>begin(B) and end(A)=end(B) | : B]A |
| begin(A)<begin(B) < end(A)<end(B) | : A/B |
| begin(B)<begin(A) < end(B)<end(A) | : B/A |

end Case

## 5. Virtual Image as bivalent interface between icons and metadata

In section 3 we introduced the story unit keyframe concept: for us it constitutes the joining element between metadata-based indexing and content–based one. Such joining is realized expanding the Virtual Image. This concept has been introduced to keyframe characterization in video segmentation and video annotation [12].

As we saw in previous section, in its original form the Virtual Image is a string of spatial relationships between objects obtained through AREM algorithm. We extend this structure providing it of Dublin Core metadata [25]. As above mentioned,] the Virtual Image is proposed as a video indexing way through the use of keyframe indexing. Then it is possibile to characterize obj not as real elements of the objects existing in the keyframe, but as elements formed by iconic image of element and metadata associated. In such way, from one side keyframe is a representative element of a video segment (shot, episodes, scenes or story unit), from the other one it is possible to index it with Virtual Image. Since in [17] the effectiveness of Virtual Image has been demonstrated in content based image indexing, we focus on the importance of introduction of metadata in the Virtual Image and in its obj elements.

In [11] there is a DDL defined using SQL-like terms for the Virtual Image and then including the metadata. We extend this concept to the streaming video providing Virtual Image of metadata at two levels. The higher level includes metadata of real image (in our case is the keyframe) from which we derived Virtual Image. Instead lower level includes metadata for the *n* objects whose Virtual Image is composed; then the $ob_j$ will be stored in a database with the AREM String and the relative metadata; such objects will be used for querying and retrieval. Then from one side Virtual Image is able to make content-based indexing (through the string of spatial relationships obtained by AREM method), from the other one it is able to index through Dublin Core metadata. Obviously it is possible to use the two methods together because Virtual Image includes both. Actually we are studying this point with many streaming video.

Then Virtual Image realizes a **biunique correspondence** between iconic content (needed by user i

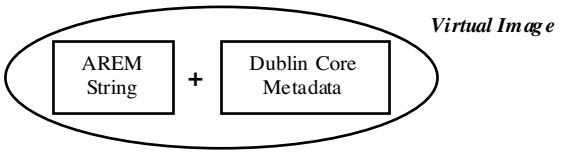



**Fig 4. The New Virtual Image Universe**


**Fig 5. Virtual Image as interface between content and metadata**

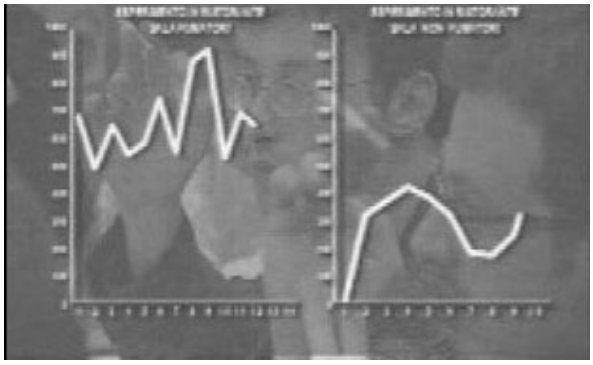


**Fig 6: Keyframe example**

the querying phase) and metadata relative to it (used by system); we schematize this concept in Fig. 5.

Let us as example a keyframe extracted from a video documentary (Fig. 6); from it we can extract the significant objects. Then we provide these objects of the Minimum Bounding Rectangle (MBR); we call the obtained image *symbolic image* (Fig. 7 ), that will be the input of the AREM algorithm.

Finally we show the Virtual Image resulting from the application of the AREM algorithm and the description through metadata in the Table 1.

**Table 1. Virtual Image and Metadata for the example keyframe**

| Keyframe | |
|---|---|
| DC.Description.text | Graphs |
| DC.Relation.IsPartOf | Scene D |
| Objects | A,B,C,D |
| AREM.X | A<B, A%C, A<D, B%D, C<B, C<D |
| AREM.Y | A=B, C/A, C/B, C=D, D/A, D/B |

As we can see in the table above we included in the Virtual Image of the keyframe the metadata and in particular:

*Description.text*, that is a little description of the keyframe (subjective information);

*Relation.IsPartOf*, that relates keyframe with the video segment (scene) or video segment aggregate (story unit) from which it is extracted (objective information).

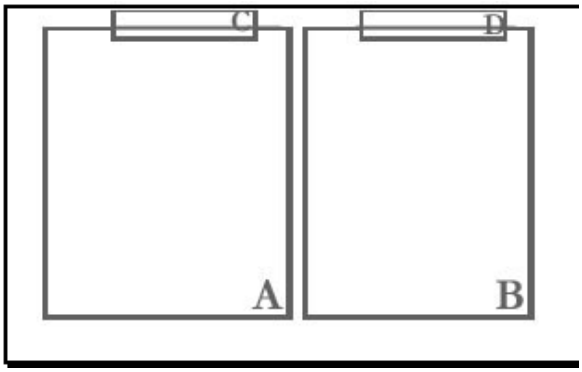



**Fig 7. Symbolic Image of the Keyframe**

Since the only metadata on the content of the keyframe (that is description.text) is a subjective information (it depends on the person assigned to database population), Virtual Image provides a more objective description of the keyframe content.

In the example keyframe there are two graphs that we havo to compare: in this case it is very important the way in which the graphs are disposed, and consequently spatial relations (between the objects) of the keyframe are important, then Virtual Image extended to metadata provides a complete description of it.

## 5. Conclusion and Future Works

In this paper we looked to integrate in a single video indexing process two different kinds of approach: the metadata based approach, based on the use of Dublin Core extentions for video streaming, and the content based one, through the use of Virtual Image. We can schematize the resulting video indexing process in Fig.8.

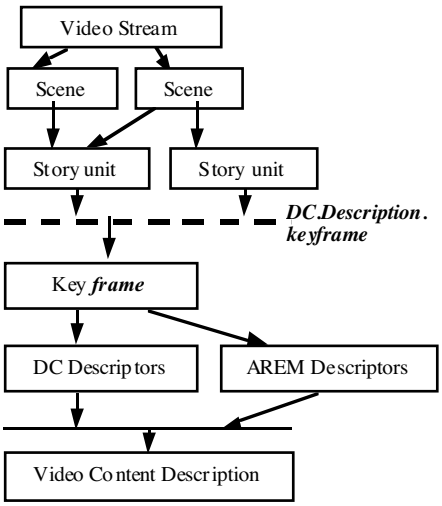



**Fig 8. The *content & metadata based* video indexing**

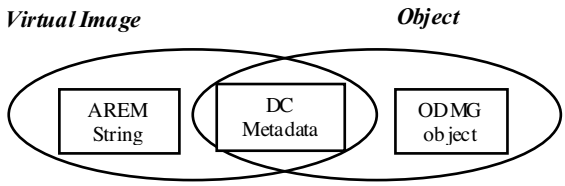



**Fig 9. Virtual Image as future Integration method**

In addition to this mechanism we are designing an integration system of the whole video indexing mechanism with an *Video Digital Repository*, based not only on an efficient video storing, scenes and frames, but also based on attributes directly derived from ODMG 3.0 standard. In such way Virtual Image will be a more important instrument for its ability to integrate the standards that are actually extending. We show this idea in Fig. 9.

## References

[1] Arndt, T. e S.K. Chang, 1989. "Image sequence compression by iconic indexing". *IEEE VL Workshop on Visual Languages*, Roma, Italy.

[2] Chang S.K., Q.Y. Shi, e C.W. Yan. "Iconic indexing by 2D-String". *IEEE Transactions on Pattern Analysis and Machine*, 9(3).

[3] Cattell R.G.G. et al. 2000. "The object data standard: ODMG 3.0" by Cattell R.G.G. and Barry D.K. Morgan Kaufmann Publisher Inc.

[4] Chang, S.K. et al. 1996. "Symbolic Projection for Image Information Retrieval and Spatial reasoning". *Signal Processing and its applications* by R. Green D. Gray and E.J. Powers Academic Press, pag. 68-70.

[5] Corridoni, J.M., A.Del Bimbo, D. Lucarella, H. Wenxue, 1996. "Multiperspective Navigation of Movies". *Journal of Visual Languages and Computing*.

[6] Davis, M., 1993. "Media Streams, an iconic visual language for video annotation". *Proceedings IEEE VL 93 Workshop on Visual Languages*. Bergen, Norway.

[7] Del Bimbo, A., 2000. "Visual Information Retrieval". Morgan Kaufmann Publishers, Inc.

[8] Del Bimbo, A., E.Vicario e D.Zingoni, 1995. "Symbolic description and visual querying of image sequences using spatio temporal logic". *IEEE Transactions on Knowledge and Data Engineering*.

[9] Flickner, M., H. Sawney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, 1995. "Query by *Image and Video Content: the QBIC System". IEEE Computer*.

[10] Freksa C., 1992. "Temporal Reasoning based on Semi-Intervals". *Artificial Intelligence*, 54 (1,2).

[11] Landi, R., L. Petraglia, G. Petraglia, 2002. "The Data Definition Language for the Virtual Image". *Proceeding of the 20th IASTED International Conference on Applied Informatics*, Feb 18-21, Innsbruck, Austria.

[12] Lee, H., A.F. Smeaton, C. Berrut, N. Murphy, S. Marlow and N.E. O'Connor, 2000. "Implementation and Analisys of Several Keyframe-Based Browsing Interfaces to Digital Video". *Research and Advanced Technology for Digital Libraries, 4th European Conference, Lisbon, Portugal. ECDL 2000.*

[13] Lee, S.Y., S.T. Lee e D.Y. Chen, 2000. "Automatic Video Summary and Description". In *Advances in Visual Information Systems*, VISUAL2000, 4th International Conference, Lyon.

[14] Lee, S.Y. and F.J. Hsu. "Spatial reasoning and similarity retrieval of images using 2D C-String Knowledge Representation". In *Pattern Recognition*, 25, 1992, 305-318.

[15] Nabil M., A.H.H. Ngu and J. Shephard, 1996. "Picture Similarity Retrieval using 2D Projection Interval Representation". *IEEE Transaction Knowledge and Data Engineering*, 8(4).

[16] Palma P. and G. Petraglia, 2002. "Video Indexing: Keyframe or Visual Icon?" In *IASTED International Conference* (submitted).

[17] Petraglia, G., M. Sebillo, M. Tucci, and G. Tortora, 2001. "Virtual images for similarity retrieval in Image databases". *IEEE Transaction on Knowledge and Data Engineering* , 13 (6).

[18] Petraglia, G., M. Sebillo, 1997. "The Virtual Image as object-relational database". *Proceeding of the 15th IASTED International Conference on Applied Informatics*, Feb 18-20, Innsbruck Austria ed M.H. Hamza IASTED Acta press, pp. 41-44.

[19] Petraglia, L., 2001. "A Federated Multidatabase System for Digital Repository on WAN". Thesis.

[20] Sawney H.S., S. Ayer e M. Gorkani, 1995. "Dominant and multiple motion for video representation". *Proceedings International Conference on Image Analysis and Processing*.

[21] Smoliar S., and H.J. Zhang, 1994. "Content-Based Video Indexing and Retrieval". In *IEEE Multimedia*.

[22] Vendrig G. and M. Worring, 2000. "Feature Driven Visualization of Video Content for Interactive Indexing". In Advances in Visual Information Systems, VISUAL2000, 4th International Conference, Lyon.

[23] http://archive.dstc.edu.au/RDU/staff/jane-hunter/ECDL3/ paper.html

[24] http://archive.dstc.edu.au/RDU/staff/jane-hunter/ECDL2/final.html

[25] www.dublincore.org

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 105-110*


# The Use of the Dublin Core in Web Annotation Programs

D. Grant Campbell
Faculty of Information and Media Studies
University of Western Ontario
gcampbel@uwo.ca

## Abstract

*This paper examines the implications of annotation programs, such as Annotea, for the development of the Dublin Core. Annotation programs enable multiple users, situated far apart, to comment on a Web-mounted document, even when they lack write access, through the use of annotation servers. Early indications suggest that the Dublin Core can significantly enhance the collaborative authoring process, especially if the full set of elements is used in a project that involves large numbers of users. However, the task of adapting DC elements and qualifiers for use in annotation threatens to increase the complexity of the scheme, and takes the Dublin Core far from its connections to traditional library cataloguing.*
**Keywords:** *Annotation programs; Semantic Web.*

## 1. Introduction

The Dublin Core is expressly committed to fostering the development of metadata description across multiple domains, and to facilitating the interoperability necessary for cross-domain resource discovery [6]. Its development has been an extended exercise in compromise, consensus-building, and dialogue among many stakeholders, including the library community and the web development community. As a result, the Dublin Core has one foot securely in the traditions of information organization; it provides a means of describing electronic resources in a way that can be mapped to traditional cataloguing standards such as the Anglo-American Cataloguing Rules, as well as frameworks for the interchange of bibliographic data, such as MARC. Its other foot rests in the emerging standards that will form the Web of the future, particularly the developments of the Semantic Web under the auspices of the World Wide Web Consortium: XML and its related standards such as Xpointer, the Resource Description Framework, and ontology creation.

These emerging standards of Web design involve not just resource discovery, but resource creation. Building on the democratizing effects of the existing World Wide Web, which has made widespread information dissemination possible to many who are shut out from the traditional publishing process, the Semantic Web seeks to broaden the Web still further by facilitating the creative process of authoring itself. Annotation programs, such as the W3C's *Annotea*, which is implemented in the W3C's Amaya browser, enable multiple users to annotate an existing document without having write access to the document's original page.

The Dublin Core stands ready to play a significant role in these annotation programs as they develop. The nature of this role, however, depends on how ambitiously the DC elements are used. And if used to its full potential in annotation, the Dublin Core could make a significant break from the document-centered cataloguing traditions that played an important part in its development.

## 2. Annotation Programs

The practice of annotation—providing commentary on information objects created at other times and usually by other people—is emerging as an important dimension of current efforts to facilitate the access and use of information on the World Wide Web. Annotation finds its most obvious use in multimedia systems, where images, sound and video can be annotated with text to facilitate retrieval. Current programs in this area range from simple captioning systems [2] to ambitious and sophisticated systems that provide multiple views of annotations in multiple formats [12].

Annotation also facilitates information retrieval in general, providing retrieval systems with additional means of eliminating spurious matches, and allowing for communication between different users of the same document store [3, 5, 9]. They also have uses in

knowledge management, by enabling organizations to tap the unofficial knowledge base of its members [10], as well as facilitating information evaluation [16].

Annotation services have always played an important part of Tim Berners-Lee's vision of a collaborative and creative Web environment:

> Imagine having servers for comments in different forums, perhaps family, school, and company. Each point and rebuttal is linked, so everyone can see at a glance the direct agreements and contradictions and the supporting evidence for each view, such that anything could be contested by the people involved. ... Again, the theme is human beings doing the thinking and machines helping it work on a larger scale, but nothing replacing wisdom in the end. [1]

Although the Web has been slower than Berners-Lee hoped at developing authoring tools, the interest in annotation programs to facilitate collaborative work is growing, as programs like Annotate!, Virtual Notes and DCRS experiment with the process of making user comments available to communities for purposes of collaborative web authoring [14, 15, 17]. The World Wide Web's contribution to this area is *Annotea*, a program that enables multiple users to provide metadata for a single pool of documents for purposes of collaborative writing and research. Three levels of use are envisioned:

- A basic level, at which annotations are used to provide commentary on a single set of documents, according to a set of categories that can be home-grown or standardized;
- A A higher level, at which both resources and annotations are bookmarked according to home-grown or standardized categories, to generate a variety of resources and metadata displays; and
- A An advanced level, at which the user-provided annotations are supplemented by metadata from other ontologies, often automatically generated. [13]

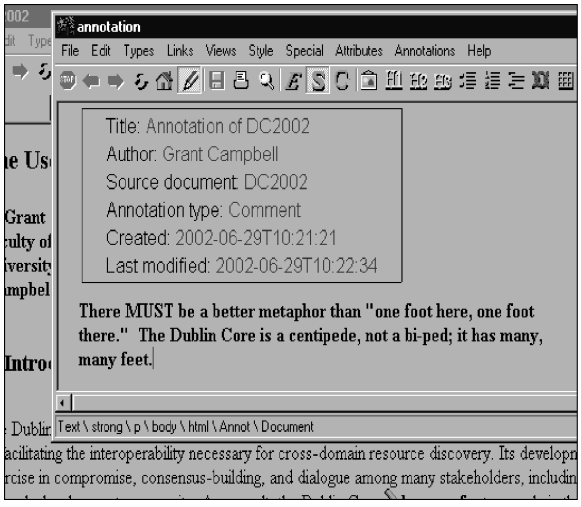



**Figure 1. Annotation using the Amaya Web Browser**

The default settings for an annotation in the W3C's Amaya browser assigns the annotation values for Title, Author, Source document, the annotation type, the date created and the date last modified (See Figure 1).

Other annotation programs, such as CREAM (CREAting Metadata), are more closely geared to the ultimate objectives of the Semantic Web, enabling either the author or another user to annotate data elements within a document with RDF metadata. Such metadata describes the data elements according to an external ontology, and clarifies their relationship with other data elements, thereby facilitating the document's use by intelligent agents [11].

Whether the task involves collaboration on the creation of a Web resource, or using an agent to assemble virtual documents in response to a specific query, the challenges facing annotation programs are formidable. Once the annotation project grows past a very few users, problems of interoperability, identification, security and timeliness present themselves. The program must be able to provide each user with the most recent annotations, and to assemble annotations efficiently, from each class of annotation, especially when classes specifically tailored to the project at hand have been created. Access and annotation rights must be limited to those authorized at each stage of the process, to preserve confidentiality.

## 3. The Dublin Core in Annotation Programs

Because of the need for interoperability, identification and access rights, the Dublin Core has a useful role to play for the annotation process. Certainly, the Dublin Core arose partly out of the recognition that metadata needed to be added at the document creation stage, and that widespread acceptance of the Core would encourage software designers to facilitate easy entry by authors [4].

Koivunen and Swick envision the Dublin Core being used to standardize the basic metadata of the annotation. Elements such as the title of the annotation, the name of the annotator, and the date created could be specified as Dublin Core elements, while other elements more specific to the annotation process could either create or use another scheme (see Figure 2).

Other programs, such as CREAM, resist the use of the Dublin Core, on the argument that the metadata, if it is to be used to facilitate the advanced semantic activity envisioned by the makers of the Semantic

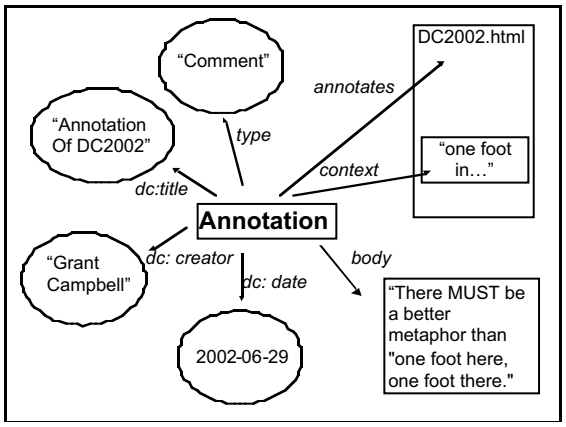



**Figure 2. Typical Annotation Scheme (Adapted from Koivunen and Swick, 2001)**

Web, must be relational. Rather than generating static notes or commentary, the metadata should be making explicit statements of relationships between class instances. According to this argument, annotation programs need to provide more than templates for entering comments, and instead provide identifiers that enable semantic relationships [11].

## 4. Enhanced Use of the Dublin Core in Annotea

The Dublin Core, however, has more relevance to annotation and collaborative creativity than either of these arguments suggest. In particular, it is possible to use DC elements more widely in the annotation process. Apart from the standard elements of Title, Creator, Contributor and Date, it is conceivable that other DC Elements could be used to express important information about the annotation for purposes of future harvesting, collocation and display.

Some DC Elements could be useful when annotation extends to a wide range of collaborators. These include:

- A Language: for use with documents being annotated across linguistic boundaries.
- A Format: for controlling the styling and display of annotations in different formats, such as HTML and XML.
- A Publisher: for annotation projects that involve a variety of individuals from multiple organizations, this element could be used to link commentators to their parent institutions.
- A Identifier: for providing the URI of the annotation.

Other elements could be used for the actual content of the annotation, as well as some of its important related information.

### 4.1 Description

The Dublin Core Reference Description defines this element as "an account of the content of the resource" [7]. Typically, it is used for abstracts, tables of contents, or some other graphical or free-text account. The text of the annotation could easily be placed in the Description element. However, such a practice does introduce an element of confusion, since the annotation functions as metadata for the original page, while the Description element serves as metadata for the annotation. Furthermore, the term "Description" does not completely apply to the spirit and purpose of annotation, which is comprises such activities as commentary, criticism, expansion, querying and references to other, related resources.

### 4.2 Type

Annotea provides a default list of annotation types, such as "advice", "change", "comment" or "question" (see Figure 3). As a description of "the nature or genre of the content of the resource" [7], the Type element could be used to classify the annotation according to a working list of categories established by the group. This would be advisable if multiple documents were being created by various subgroups that would later need to be joined together.

The question then arises: to what degree should the Dublin Core provide qualifiers to the "Type" element to facilitate annotation activities? Certain activities, such as comment, change and question might be considered universal, and worth defining at the level of the metadata set for interoperability purposes. Others may well be defined by a specific group for its own purposes.

### 4.3 Coverage

While coverage is usually conceived in temporal or geographical terms, it could also be used in a collaborative context to indicate:

- The range of annotation. In this way, aggressive and far-reaching commentary, appropriate to the initial brainstorming stages of a project, could be separated from the grammatical, stylistic and technical annotations appropriate for the proofreading stages.
- The area of the document covered; annotations of one section, such as the introduction could then be separated from those directed at others, such as the bibliography, or FAQ page.
- The stage of consultation: annotations on an annual report, for instance, could be classed according to those provided by the original team of authors, those provided by the organization as a whole, those provided by government or other external officials, and those provided by the general public.

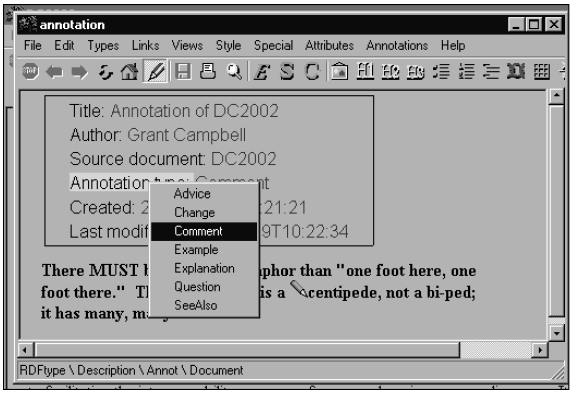



**Figure 3. Annotation Types in the Amaya Browser**

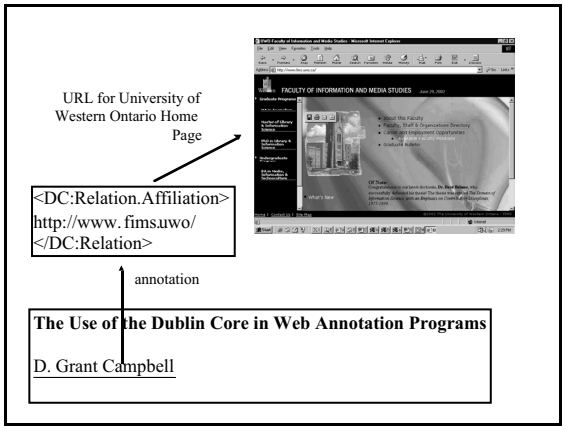



**Figure 4. Using the Relation Element to Express a Relationship**

### 4.4 Rights

This element could be used to administer access to the various annotations. In the case of a document on a sensitive subject which reflected the collective thought of a committee or other administrative body, such administration would preserve the privacy of those involved in the original deliberations, after the point when the committee's decision is made.

### 4.5 Relation

This element references "a related resource" [7], and as such may be the most significant element for the Dublin Core as it adapts both to collaborative authoring and to the Semantic Web. At present, the established DC refinements support such relationships as versioning, replacement, and part relations. In a simple annotation process, this element could be used for the URI of the original document. With an expanded list of refinements, this element could also be used for relational metadata, thereby embedding semantic relationships that could be used for sophisticated machine processing. In the Figure 4, for instance, an annotation of the author's name creates a link to the author's home faculty. Such a link helps to identify the author as the "Grant Campbell" who teaches for the Faculty of Information and Media Studies, and disambiguates him from others with the same name.

By using the Dublin Core elements to their full potential, therefore, an annotation could look something like this:

<?xml version = "1.0"?>
<RDF
xmlns = "http://www.w3.org/TR/1999/REC-rdf-syntax-19990222#"
xmlns:DC = "http://metadata.net/dstc/DC-10-EN/#">
<Description xml:lang="en">
<DC:Title>**Annotation of DC2002**</DC:Title>
<DC:Creator>**Campbell, Grant**</DC:Creator>
<DC:Description>
**There MUST be a better metaphor than "one foot here, one foot there."**
</DC:Description>
<DC:Publisher> **Faculty of Information and Media Studies, University of Western Ontario**
</DC:Publisher>
<DC:Date DC:Scheme="ISO8601">**2002-06-29T10:21:21**</DC:Date>
<DC:Type>**Commentary**</DC:Type>
<DC:Format DC:Scheme="IMT">**HTML** </DC:Format>
<DC:Identifier DC:Scheme="URI"> **http://instruct.uwo.ca/fim-lis/502/** </DC:Identifier>
<DC:Language DC:Scheme="RFC1766"> **EN**</DC:Language>
<DC:Relation.Annotates>**http://instruct.uwo.ca/fim-lis/502/dc2002.htm**</DC:Relation>
<DC:Rights>**For use within the authoring circle only**.</DC:Rights>
</Description>
</RDF>

## 5. Consequences and Conclusions

Widespread use of the Dublin Core in annotation programs could be highly beneficial to the annotation process. Use of DC elements and qualifiers could simplify the collation and ordering of annotations through standardized versions of dates and formats, and many of the elements could be used, as shown, to do justice to the rich complexity of the collaborative process involved in conceiving, creating, revising and disseminating documents. The annotation process, however, revives certain tensions that have always plagued traditional methods of document description and retrieval, while simultaneously breaking down many of the distinctions that have given these methods their coherence. Making annotation a recognized and important part of the Dublin

Core's purpose could carry profound consequences for DC activities and developments.

### 5.1 Detail vs. Speed

The use of DC elements for extensive support of annotation threatens to revive the structuralist/ minimalist debate that has plagued the Dublin Core for years. While many of the elements can be meaningfully adapted, some, such as the "Relation" and "Type" elements, will need further qualification. And annotation users will be sorely tempted to "smarten up" the Dublin Core to do justice to the subtleties and rich demands of the collaborative process, just as the cataloguing community has introduced qualifiers and refinements to enhance interoperability with MARC records. If annotation programs continue to proliferate, developers may well decide to limit use of the Dublin Core to the few elements that can be simply and unambiguously applied, choosing to extend it with new schemas and alternate schemes as desired. Extensibility, after all, is a fundamental principle of a metadata set that strives to be a core, not a comprehensive descriptive code [8].

### 5.2 Document vs Data

Beneath this revival of the structuralist/ minimalist controversy lies an even more interesting trend. With annotation programs, the Dublin Core is finally moving into a new bibliographic universe: one that we've always been aware of, but have been only fitfully able to inhabit. Unlike the traditional bibliographic universe, which consists of physical documents which are aggregated by catalogues into meaningful units such as editions, series and works, this new universe is highly granular, and breaks documents down before aggregating the individual data elements. This universe, which owes as much to computer science and database design as to traditional library science, is a universe of "entities" and "relationships": terms which are deliberately amorphous and vague, and whose meanings are assigned locally within specific communities and domains, linked tenuously together by ontologies.

This universe of data given local context and assembled across domains through ontologies is, of course, the universe of the Semantic Web. And annotation programs in their various uses break down distinctions that have traditionally prevailed in document organization and description. Annotation reduces the gulf between textual and non-textual information sources by providing a textual dimension to multimedia artifacts. It breaks down the distinction between official and non-official publication, by facilitating unofficial comment on official documents, thereby mobilizing the vast amount of hidden knowledge available in a community or a workforce. And annotation collapses the distinction between information retrieval and information evaluation, by bringing the community into the retrieval process, and providing additional means by which information can be evaluated at the retrieval stage.

As the Dublin Core moves towards the envisioned world of the Semantic Web, it stands to benefit from the foresight of its initial founders, who, in 1995, chose to address the problem of describing digital objects in general, rather than specifying electronic "documents", "books", or "articles". With the rise of annotation programs, we can see the movement of the Dublin Core away from the "document", whether it is a resource in CORC that must be represented either in DC or MARC, or a resource harvested through an OAI harvesting system. The Dublin Core is breaking through that document layer, and is now describing and addressing discrete data units that can be detached, collated and assembled in fresh and dynamic ways.

## References

[1] Berners-Lee, T. and Fischetti, M. *Weaving the Web: the Original Design and Ultimate Destiny of the World Wide Web*. New York: HarperBusiness, 2000.

[2] Bertino, E., B. Catania, and G.P. Zarri. "A Conceptual Annotation Approach to Indexing in a Web-Based Information System". *International Conference on Advance Issues of E-Commerce and Web-Based Information Systems*. WECWIS, 1999, pp. 160-165.

[3] Bremer, J.M. and M. Gertz. "Web Data Indexing through Semantic-Carrying Annotations". *Proceedings of the 11th Annual Workshop on Research Issues in Data Engineering*, 2001, pp. 69-76.

[4] Caplan, P. and R. Guenther. "Metadata for Internet Resources: The Dublin Core Metadata Elements Set and its Mapping to USMARC". *Cataloging & Classification Quarterly*, vol. 22, pp. 43-58, 1996.

[5] Czejdo, B., J. Dinsmore, C.H. Hwang, R. Miller, and M. Rusinkiewicz. "Automatic Generation of Ontology Based Annotations in XML and their Use in Retrieval Systems". *Proceedings of the First International Conference on Web Information Systems Engineering*, 2000, pp. 296-300.

[6] Dekkers, M. and S. Weibel, S. "Dublin Core Metadata Initiative progress report and workplan for 2002". [online] *D-Lib Magazine*, 8(3). Available from: http://www.dlib.org/dlib/february02/weibel/02weibel.html [June 29, 2002].

[7] Dublin Core Metadata Initiative, *Dublin Core Metadata Element Set*, Version 1.1: Reference description. [online] Available from:

http://dublincore.org/documents/dces/ [June 29, 2002].

[8] Duval, E., W. Hodgin, S. Sutton, and S. Weibel, "Metadata principles and practicalities". [online] *D-Lib Magazine*, 8(4). Available from: http://www.dlib.org/dlib/april02/weibel/04/ weibel.html [June 29, 2002].

[9] Ginsburg, M. "Annotate!: A Tool for Collaborative Information Retrieval". *Proceedings of the Seventh IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*, 1998, pp. 75-80.

[10] Ginsburg, M. and A. Kambil. "Annotate: A Web-based Knowledge Management Support System". *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences*, 1999.

[11] Handsuch, S. and S. Staab, "Authoring and annotation of Web pages in CREAM." [online] In *Proceedings of the 11th International World Wide Web Conference, Honolulu, May 7-11, 2002*. ACM. Available from: http://www2002.org/CDROM/refereed/506/index.html [June 29, 2002].

[12] Hirotsu, T., T. Takada, S. Aoyagi, K. Sato and T. Sugawara. "Cmew/U: A Multimedia Web Annotation Sharing System", TENCON 99: *Proceedings of the IEEE Region 10 Conference*. Vol 1, 1999, pp. 356-359.

[13] Koivunen, M. and R. Swick. "Metadata based annotation infrastructure offers flexibility and extensibility for collaborative applications and beyond." [online] In *K-CAP 2001: First International Conference on Knowledge Capture, Victoria 21-23 October*. Available from: http://sern.ucalgary.ca/ksi/K-CAP/K-CAP2001/ [June 29, 2002].

[14] Koch, S. and G. Schneider. "Implementation of an Annotation Service on the WWW: Virtual Notes." *Proceedings of the 8th Euromicro Workshop on Parallel and Distributed Processing*, 1999, pp. 92-98.

[15] Richy, H. and G. Lorette. "On-line Correction of Web Pages." *Proceedings of the 5th International Conference on Document Analysis and Recognition*, 1999, pp. 581-584.

[16] Sannomiva, T., T. Amagasa, M. Yokishawa, and S. Uemura. "A Framework for Sharing Personal Annotations on Web Resources Using XML." *Proceedings of the 2001 Workshop on Information Technology and Virtual Enterprises*, 2001, pp. 40-48.

[17] Tsai, S.R., J. Chen, and M. Koo. "A Document Workspace for Collaboration and Annotation Based on XML Technology", *Proceedings of the 2000 International Symposium on Multimedia Software Engineering*, 2000, pp. 165-172.

Paper Session 3

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 113-123*


# A Comprehensive Framework for Building Multilingual Domain Ontologies: Creating a Prototype Biosecurity Ontology

Boris Lauser, Tanja Wildemann, Allison Poulos,
Frehiwot Fisseha, Johannes Keizer, Stephen Katz
Food and Agriculture Organization of the UN, GILW
Rome, Italy
{boris.lauser, tanja.wildemann, allison.poulos,
frehiwot.fisseha, johannes.keizer, stephen.katz}@fao.org
http://www.fao.org

## Abstract

*This paper presents our ongoing work in establishing a multilingual domain ontology for a biosecurity portal. As a prototypical approach, this project is embedded into the bigger context of the Agricultural Ontology Service (AOS) project of the Food and Agriculture Organization (FAO) of the UN. The AOS will act as a reference tool for ontology creation assistance and herewith enable the transfer of the agricultural domain towards the Semantic Web. The paper focuses on introducing a comprehensive, reusable framework for the process of semi-automatically supported ontology evolvement, which aims to be used in follow-up projects and can eventually be applied to any other domain. Within the multinational context of the FAO, multilingual aspects play a crucial role and therefore an extendable layered ontology modelling approach will be described within the framework. The paper will present the project milestones achieved so far: the creation of a core ontology, the semiautomatic extension of this ontology using a heuristic toolset, and the representation of the resulting ontology in a multilingual web portal. The reader will be provided with a practical example for the creation of a specific domain ontology, which can be applied to any possible domain. Future projects, including automatic text classification, and ontology facilitated search opportunities, will be addressed at the end of the paper.*

**Keywords:** *Ontology, Semantic Web, Ontology creation, Ontology Engineering Framework, Ontology Learning, Multilingual Ontology, Biosecurity, Food Safety, Animal Health, Plant Health.*

## 1. Introduction

### 1.1 Motivation and subject domain

The management of large amounts of information and knowledge is of ever increasing importance in today's large organizations. With the ongoing ease of supplying information online, especially in corporate intranets and knowledge bases, finding the right information becomes an increasingly difficult task. Today's search tools perform rather poorly in the sense that information access is mostly based on keyword searching or even mere browsing of topic areas. This unfocused approach often leads to undesired results. The following example illustrates the problem more clearly:

One might, for example, want to find out which organization established the Agreement of Agriculture. A simple search for "establish Agreement of Agriculture" might result in a huge list of documents containing these words, but actually none of them containing the desired result: WTO or World Trade Organization. The problem becomes even worse, if the result searched for only appears in a foreign language document. Figure 1 shows an extract of an ontology, which could solve this problem. The grey ellipses represent generic concepts, whereas the white ones represent specific instances of these concepts. The two concepts shown here are interlinked by a relationship. The ontology enabled search application would first identify "Agreement of Agriculture" as a "standard" and would then detect the relationship "establish" to "international organization"

and its instances, and hence solve the problem by extending the search query. Furthermore, it could provide added value by detecting other relationships that provide the user with more possibilities, for example standards of other organizations could be presented.

This example shows how ontologies can help to improve the management of information. Semantically annotated documents, i.e. documents which are indexed with ontological terms and concepts instead of simple keywords, provide several advantages. First, the ontological abstraction provides robustness against changes in the document. In the above example, the document content might change using the term 'Agricultural Agreement' instead of 'Agreement of Agriculture'. However, since the document has been annotated with the ontological semantics, this will not affect the search results. Second, since the ontology used for annotating the document is domain specific, the semantic meanings and interpretations of keywords are bound to that domain and therefore the retrieval is likely to be more efficient. A term can have several meanings in different domains. By first mapping the keyword to its semantic representation in a specific ontology and using the ontology's linked knowledge structure, a much more focused search approach can be taken. Third, document specific representations no longer affect the search. This is extremely important in the case of multilingual representations. Keywords of several languages are mapped to the same concept in an ontology and are therefore given the same meaning. Multilingual search portals can be established to produce the same results, no matter which language is used for retrieval.

Another important issue of knowledge management, especially with regards to document metadata and indexing, is the classification of documents. Presently, this is carried out by subject specialists in a time consuming process. With today's vast amount of available information on the WWW, automatic support is needed to efficiently manage this task. Ontologies play a critical role in supporting the machine readable semantics needed to facilitate automation.

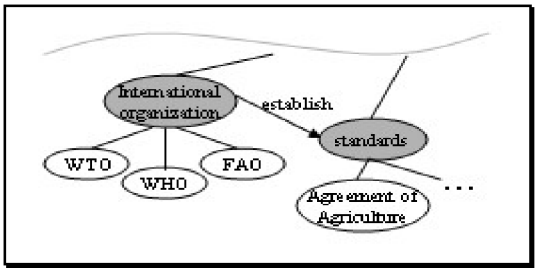



**Figure 1. Ontology example, excerpt**

Before such powerful Semantic Web[1] applications can be built and used within certain domains of knowledge, the basic requirement, a machine readable vocabulary represented by a domain ontology has to be established. The creation of ontologies is a time consuming task and often carried out in an ad-hoc manner. Only few methodologies exist, and even less automated tool support is available. Constituting the knowledge base for future Semantic Web applications, domain ontologies have to be created continuously in all possible areas and communities. The need for a reusable methodology is evident. This paper outlines a comprehensive, reusable framework for semi-automatically-aided building of domain ontologies. A prototype project is used for the application of this computer-aided framework, which provides the reader with a practical, methodological ontology engineering approach.

The domain that serves for creating the prototype ontology is the Biosecurity Portal on food safety, animal and plant health. The portal is an access point for official national and international information relating to biosecurity, the risks associated with agriculture (including fisheries and forestry), and food production. Many countries are still struggling with rapid advances in technology and often lack access to basic information on food safety, animal health and plant health. However, access to this information is of paramount importance for countries to protect health, agriculture and the environment.

One of the goals of the portal is to serve as an electronic information exchange mechanism for the addressed community and therefore to ensure efficient and effective information retrieval. The extension of its knowledge base to information available on various other sources in the WWW can highly support the purpose of the portal. Serving an international community, the information must be retrievable in various languages. The domain is multidisciplinary across three different, but related subject areas. The motivation to create a commonly agreed on, formally specified vocabulary in form of domain ontologies becomes evident

### 1.2 Overview of the approach

The presented project introduces a comprehensive framework for building a domain-specific ontology. The approach combines classical methodologies for human-based ontology engineering with semiautomatic support of a heuristic toolset. Actually, two methods for ontology acquisition are applied in order to create the domain ontology. The first is to create a small domain-specific core ontology from scratch and then apply a focused web crawler to this ontology in order to retrieve domain related web pages and interesting domain terms for extending this base

[1] Refer to (Palmer 2001) for an introduction to the Semantic Web.

ontology. The second acquisition approach takes a well-established thesaurus as a basic vocabulary reference set and converts it to an ontology representation. Then a domain specific and a general corpus of texts are used in order to remove concepts that are not descriptive for the domain. The heuristic used here is, that domain specific concepts are more frequent in the domain-specific text corpus. A side product of this removal step is again a list of frequent terms, which can eventually enhance the ontology (see Volz 2000 for more details on this approach). The results of these steps are assessed to assemble the final domain specific ontology, which is now accessible through a multilingual web portal.

### 1.3 Outline

The next section provides a brief introduction to the larger framework the prototype project is embedded in. In Section 3 a proposed layered multilingual ontology model is introduced. It sets the basis for the methodological framework, which is discussed in detail in Section 4. All steps of the prototype project are then presented in Section 5 and currently available results are shown. Finally, Section 6 gives an outlook on further work and opportunities that this project enables.

## 2. The project framework: FAO and the AOS

The Food and Agriculture Organization (FAO) of the United Nations (UN) is committed to helping combat and eradicate world hunger. Information dissemination is an important and necessary tool in furthering this cause, and we need to provide consistent, usable access to information for the community of people doing this very work. The wide recognition of FAO as a neutral international centre of excellence in agriculture positions it perfectly to lead in the growth and improvement of knowledge representation systems in the agricultural domain.

Above discussed Semantic Web applications could contribute to this mission. The need for improved information management mechanisms within the various knowledge domains of this organization is therefore evident.

The Agricultural Ontology Service (AOS) Project evolved from this motivation and has been initiated to act as a reference tool for ontology creation to enable the transfer of the agricultural information domain towards the Semantic Web. The goals of the AOS are to increase the efficiency and consistency of describing and relating multilingual agricultural resources, to decrease the random nature and increase the functionality for accessing these resources and to enable sharing of common descriptions, definitions and relations within the agricultural community. To achieve these goals the AOS assists community partners in the creation of ontologies and related activities. The project, which will be presented in this document, serves as a prototype within the AOS framework and shall serve as a reference to further activities. A comprehensive and reusable methodology, which can be applied to any other domain, is to be evaluated by this prototype. A multilingual, extendable model for the representation of domain ontologies builds the core baseline of this methodology and will be presented in the following section.

## 3. The ontology: Modelling and representation

In the context of the AOS, an ontology is a system of terms, the definition of these terms and the specification of relationships between the terms. It extends the approach of classical thesauri by providing the opportunity of creating an infinite number of different semantic relationships. For an overview about different types of ontologies, refer to (Guarino 1998). The following gives a detailed description of the modelling approach used for our representation of the prototype ontology:

Semantic robustness towards representational changes, as well as multilingualism, are crucial for the development of this domain ontology (see section 1.1). Therefore, we distinguish between terms, and the concepts these terms represent. Whereas terms might change, and are different in each language, the semantic meaning and interpretation of the terms' abstract concept stays the same[2]. In the presented modelling approach, a concept's term representations are called Lexical Entries. These Lexical Entries are limitless and may be characterized as labels, synonyms or word stems. Furthermore, each Lexical Entry has at least two attributes: the concept it refers to and its language. Lastly, relationships between concepts can be established, annotated by the same lexical entries. This approach can be described as a two layered model, in which the semantic layer of the ontology is totally independent from its representation layer and hence, robustness against changes can be achieved.

Ontologies can be represented in different representation languages. (Palmer 2001) gives a brief overview about these languages and provides further information. RDFS[3], the language that was chosen to

[2] This holds in most cases. There are however cases, where a concept does exist in one culture, even though there is not adequate concept in another one. This is however more evident in humanity domains, since concepts there are richer and less well defined. The project environment here is rather technical and hence chances for this can be neglected.

[3] http://www.w3.org/TR/rdf-schema/#intro.

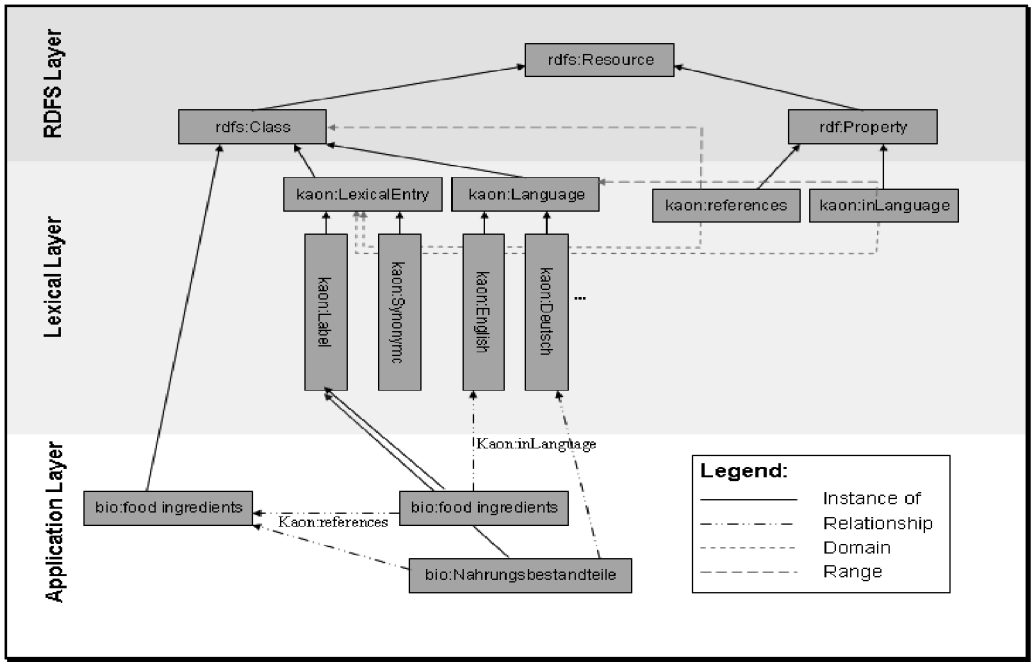



**Figure 2. Layered RDFS model multilingual ontology representation**

be used within the AOS framework, is used to define vocabularies of resources and relationships amongst them. Resources can be documents, web pages or parts of them, basically anything, which can be referenced by a URI[4]. RDFS provides a basic set of modelling primitives, which can be easily extended by users to include domain specific semantics in terms of relationships among concepts. Furthermore RDFS models are exchanged via XML and therefore provide interoperability between communities. Although still under development, RDFS evolves to serve as a standard representation in the context of the Semantic Web. For a detailed discussion about modelling ontologies in RDFS, refer to (Staab et al. 2000a).

Figure 2 gives an overview of the above-discussed layered modelling approach in RDFS. The top layer represents an extract of the basic layer provided by the RDFS language. The lexical layer creates the needed abstraction of lexical and language representation from conceptual domain semantics. The lowest layer finally constitutes the domain. The most generic class in RDFS is rdfs:Resource[5], from which every other class is derived An rdfs:Class can be instantiated to define domain specific concepts. Lexical Entries are separate classes which can be instantiated and attached to concepts using the properties kaon:references and kaon:inLanguage. Each property has a domain and a range, which determine the source and the target of the relationship respectively. In that way, an infinite number of lexical entries can be instantiated and related to domain concepts and different languages. If a representation of a concept in terms of its lexical entry changes, the semantics of the ontology are not affected, since it still refers to the same concept. Furthermore, additional domain properties can be derived from rdf:Property in the application layer to relate the domain concepts and build the semantic network.

This generic, multilingual ontology model establishes the basis for our engineering methodology framework, which will be presented in the following section.

## 4. The methodological framework

Until now, few domain-independent methodological approaches have been reported for building ontologies. Most of these are mainly overall lifecycle models providing a more generic framework for the ontology creation process, but giving little support for the actual task of building the ontology. A comparative study of ontology building methodologies from scratch can be found in (Fernandez 1999). The METHONTOLOGY methodology, as described in (Fernandez et al. 1998) fits our project approach best, since it proposes an evolving prototyping life cycle composed of development oriented activities (requirements specification, conceptualization of domain knowledge, formalization of the conceptual

[4] Uniform Resource Identifier. See also http://www.w3.org/Addressing.

[5] The prefixes rdfs: , rdf: , kaon: , bio: represent XML namespaces and are to uniquely identify each resource. Refer to (RDFSchema 2002) to learn more about RDFS and namespaces.

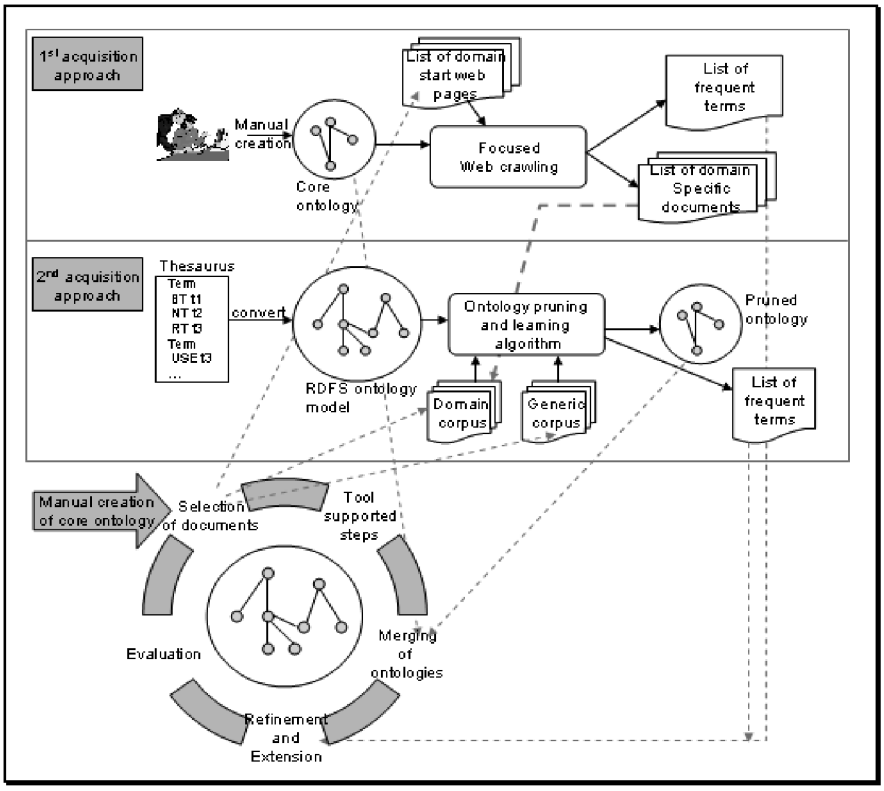



**Figure 3. Comprehensive framework for creation of domain ontologies**

model in a formal language, implementation of the formal model and maintenance of implemented ontologies), support oriented activities (knowledge acquisition, documentation, evaluation, integration of other ontologies) and project management activities. Since this has been done elsewhere, the framework presented in this paper will not propose another life cycle model. Rather, it will depict the development oriented activities within the above methodology and provide a more specific methodology for this part. More specific methodologies, especially for supporting the creation process sparsely exist so far. (Guarino et al.) provide a set of methodologies for ontology-driven conceptual analysis. An overview of these methodologies can be accessed through his web site. The methodology presented here focuses on the actual acquisition and development part of the ontology and describes a comprehensive, reusable and semi automatically-supported framework, which can be embedded in other lifecycle models. Figure 3 shows a graphical overview of the overall framework.

The domain ontology is built using two different knowledge acquisition approaches, which will be described in detail in the following sections. The top of the picture describes these two paths. In the lower part of the picture the cyclic evolvement of the domain ontology to be built is shown. The grey dashed arrows show how outputs of certain processes steps are used as inputs of other steps. Section 5, where the application of this framework to the biosecurity prototype is presented, will present each single process step and its application to the prototype project.

## 5. The biosecurity ontology project

### 5.1 Acquisition approach 1: Creation of the core ontology

In the first acquisition approach, a small core ontology with the most important domain concepts and their relationships is created from scratch. This stage is basically comprised of the first three steps of the METHONTOLOGY development activities (as described in section 4):

First the goal of the ontology is specified (as outlined in section 1.1 and in section 2). In a second step, subject specialists accomplish the conceptualization of the core model. The Codex Alimentarius, which serves as a reference for food standards in food safety biosecurity, has been chosen here for extracting basic domain concepts. In further brainstorming sessions, relationships between the chosen concepts and additional concepts are created. The concepts and relationships are further assessed using criteria including clarity, ambiguity, unity and rigidity. A detailed discussion of criteria for ontology-driven conceptual analysis is given in (Welty 2001).

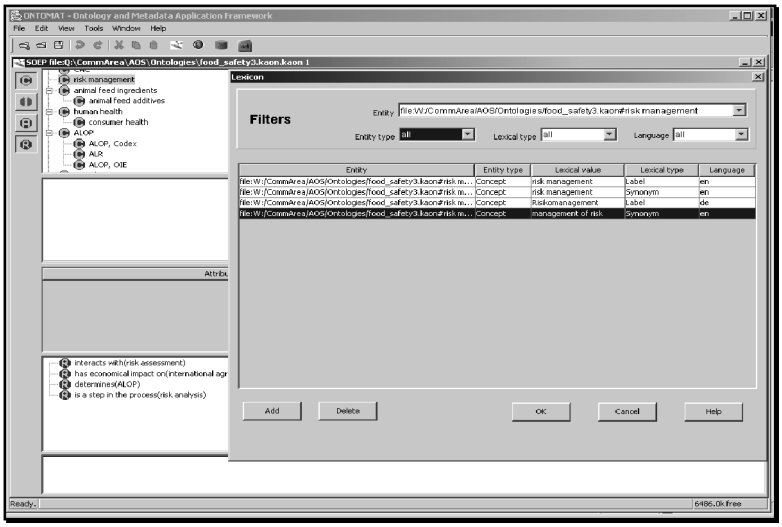



**Figure 4. Screenshot of the ontology editor SOEP**

In the biosecurity project, this initial step created a core ontology with 67 concepts and 91 relationships connecting these concepts, equalling an average rate of 1.36 relationships per concept.

Finally the developed core ontology is formalized in the formal RDFS language. This can be accomplished using the RDFS compatible ontology editor SOEP[6] of the KAON[7] tool environment. The editor has an easy-to-use graphical user interface, which allows the creation of the concepts, their relationships and their lexical entries. Figure 4 shows a screenshot of the resulting core ontology in the editor. On the upper left, concepts and their hierarchical subclass relations are shown. On the lower left, one can see the domain specific relationships between a marked concept and other concepts. The additional window on the right side shows the lexical layer of the ontology. This clearly illustrates that the entities (in this case the concept 'risk management') are represented uniquely by a URI, therefore unambiguous, and a concepts lexical entries are all independently associated with this URI.

In the following acquisition stage, the core ontology is fed into a Focused Web Crawler, another tool of the KAON environment. The Crawler takes a set of start URLs and domain ontology. It then crawls the web in search of other domain specific documents based on a large set of user specified parameters. The outcome this process creates consists of a rated list of found domain specific documents and links as well as a list of most frequent terms found on these documents. A list with 264 domain-relevant web pages and a list with 36 frequent terms have been output by the crawler in our prototype project. The list of keywords can later be used to extend the core ontology. The document list can be used as input in the second ontology acquisition approach, which will be described in the following section.

### 5.2 Acquisition approach 2: Deriving a domain ontology from a thesaurus

The second approach towards ontology acquisition takes a well-established thesaurus as starting point. Here, AGROVOC[8], a multilingual agricultural thesaurus consisting of almost 30,000 keywords developed by the FAO, is assumed to contain domain descriptors. A thesaurus like AGROVOC consists of descriptive keywords linked by a basic set of relationships. The keywords are descriptive in terms of the domain in which they are used. The relationships may either describe a hierarchical relation or an inter-hierarchical relation. For example, 'Broader Term' and 'Narrower Term' are used for the former and 'Related Term' and 'Use' for the latter. The 'Use' relationship indicates that another term should be used for description instead of this one.

The process begins by representing the thesaurus in an adequate format, where an ontology can be derived from. As discussed above, RDFS is chosen as the representation language. Then, as done in the biosecurity ontology, all terms of the thesaurus are converted to classes (concepts)[9]. The Broader and Narrower Term relationships are used to form the hierarchical class-subclass structure, which constitutes the basic taxonomy of the ontology. Finally the Related Term and Use relationships are represented as properties of the classes and form an initial set of non-hierarchical relationships. This approach extends the basic RDFS language by creating new, layered meta-properties, which can be instantiated in

[6] Simple Ontology and Metadata Editor Plugin.
[7] Karlsruhe Ontology and Semantic Web Tool Suite.
[8] http://www.fao.org/agrovoc
[9] In this paper, classes and concepts are synonymous, where class refers to the RDFS representation of the concept in an ontology.

```
...
<rdf:Property rdf:ID="rt">
        <rdfs:domain rdf:resource="&kaon;Root"/>
        <rdfs:range rdf:resource="&kaon;Root"/>
</rdf:Property>
....
<rdfs:Class rdf:ID="7">
        <rdfs:subClassOf rdf:resource="#1172" />
        <rt rdf:resource="#3471" />
</rdfs:Class>
...
```

**Figure 5. Extract of RDFS modelling of the AGROVOC thesaurus, using meta properties**

the domain classes. The modelling is done analogously to the above described language layer. Figure 5 gives an example representation of the Related Term definition and a class using this relationship in RDFS. Here the concept with the identifier 7 is a sub class of concept 1172 and is related to the concept with the identifier 3471. Lexical labels for representation in different languages are attached to these concepts and relations as discussed before.

The converted thesaurus still has to be trimmed to the specific domain. An ontology pruner is used to accomplish this task. In order to prune the thesaurus structure to extract a domain-specific ontological structure, two sets of documents are needed: a domain specific set, descriptive for the domain of the goal ontology to be built, and a generic set, containing a representative set of generic, unspecific terms. This step can partly be done before the tool supported steps and therefore appears on top of the cyclic process in Figure 3. The domain documents have to be carefully chosen by subject specialists. The output of the process obviously correlates with the descriptiveness, preciseness and richness (in means of specific domain term usage) of the domain document set. The document list, which is the outcome of the web crawling process, can serve as a good source. Publicly available reference corpora and newspaper archives serve as sources for the generic corpus. In addition, sets of related, but different, subject domains may also be used. This could increase the chances of retrieving only very specific concepts, since the terms' frequencies of the domain corpus are measured against those of the generic corpus. However, the whole process is a highly heuristic approach and further experiments are needed to establish a significant document set quality measure.

In our case, a set of six domain specific documents (mainly excerpts of the Codex Alimentarius, as well as documents about food safety and risk assessment) has been chosen and another eight documents have been taken from the list of the crawling process. The generic document set has been compiled using news web pages, as well as pages from the animal feed domain, another research area within the FAO.

In order to prune domain unspecific concepts, concept frequencies are determined from both domain-specific and generic documents. All concept frequencies are propagated along the taxonomy to their super concepts by summing the frequencies of sub concepts. The frequencies of the concepts in the domain corpus are then compared with those of the same concepts in the generic corpus using pruning criteria. Only the concepts, which are significantly more frequent in the domain corpus, remain in the ontology, the others are discarded. Moreover the frequencies of all terms occurring in the domain documents can be compared against all the terms that occur in the generic corpus resulting in a list of terms, likely to be significant for the domain corpus. Refer to (Volz 2000) for a detailed discussion on ontology acquisition using text mining procedures and to (Kietz 2000) for a similar application of extracting a domain ontology.

The result of the second ontology acquisition approach is a pruned ontological structure derived from the original thesaurus, containing only the domain specific terms. It also produces a list of likely domain-specific terms, which can serve as possible candidates for the ontology refinement process.

Here, an ontological structure with 504 concepts could be extracted from the AGROVOC thesaurus with a taxonomic depth of five. A list of 1632 frequent terms has been produced from the domain document set.

### 5.3 Ontology merging

The above acquisition steps have created two ontologies, the manually created core ontology and the derived ontology, using thesaurus terms. These have to be assembled into a single ontology. Ontology merging is still more of an art than a well-defined and established process. (Gangemi et al.) describe a methodology for ontology merging and integration in the Fishery Domain. Besides the editor environment, computer support for this process is not available and therefore needs extensive subject specialist assessment.

From the pruned ontological structure of the AGROVOC thesaurus, 23 concepts and 13 instances have been extracted to extend the core ontology in our case. Hence, almost 10% of the automatically extracted knowledge could be used in the first instance. More terms might serve as candidates in further refinement steps.

### 5.4 Ontology Refinements and Extension

The second result produced by the acquisition steps is a list with frequent domain terms serving as possible candidate concepts or relationships for extending the ontology. These terms have to be assessed by subject specialists and checked for relevance to the ontology. The same principles and methodologies, as in the creation process of the core ontology, apply to this session. In our case, 12 concepts were directly taken from the lists of frequent

keywords to extend the ontology. A set of 12 new unique relationships has been defined, resulting in 92 relations interlinking and integrating the newly created concepts. These have been applied to assemble the final prototype ontology consisting of 102 concepts, 12 instances and 183 relationships among the concepts. This corresponds to an average rate of 1.79 relationships per concept, representing a higher density than in the core ontology.

The resulting ontology is now subject to more extensive evaluation and testing by a broader audience. The presentation of the ontology in a multilingual portal, which will be presented in the next section, can help in the evaluation process. However, extensive testing and evaluation cannot be done effectively until real applications utilize the semantic power of the ontology. This will be addressed in the last section, where an outlook on further work and future uses will be given.

### 5.5 Presentation in Multilingual Portal

The domain ontology can be extended to represent the concepts in multiple languages. The translation process has to be done manually, since current translation tools show rather inferior performance and are also quite unlikely to be applicable to specific domains like the biosecurity portal. With our ontology model introduced in Section 3, this task can easily be achieved by simply attaching further lexical entries to the concepts of the newly created ontology. In the project presented here, this step has been omitted since it is not of importance to prototype versions. Finally, KAON PORTAL, a web-based portal to present RDFS based ontologies, can be used to present the ontology, making it available and browseable to the target community. Figure 6 shows a screenshot of the top concept layer of the prototype Biosecurity Ontology. The display can be switched to different languages, including Arabic and Chinese.

This portal could now be extended to actually link to a domain document base and the ontology could be used to perform semantically extended search opportunities.

## 6. Outlook: Future uses of the ontology (implementation of the semantic web)

In this section, an outlook on future work within this project and follow-up projects in context of the AOS framework will be given. As previously discussed, a domain ontology, which can be developed applying the above framework, only sets the basis for efficient information management and retrieval. Applications, using this background knowledge are still rare and further investigation is required. This section sketches a likely scenario for ontology use in the discussed domain and outlines some already existing sample applications and their possible implications for the AOS project.

### 6.1. Facilitation of better search and information retrieval

Using the ontology to extend currently performed keyword search, is the most direct application. Ontology based support could be given at two stages of the search query process: before the actual execution of the query and/or after retrieval of the results. Figure 7 shows these two semantically enhanced search features. The left side shows a scenario, in which the ontology assists the user by providing an easy way to extend or refine her search. The ontology enabled search application processes on the initial search term. It then queries the ontology to retrieve the

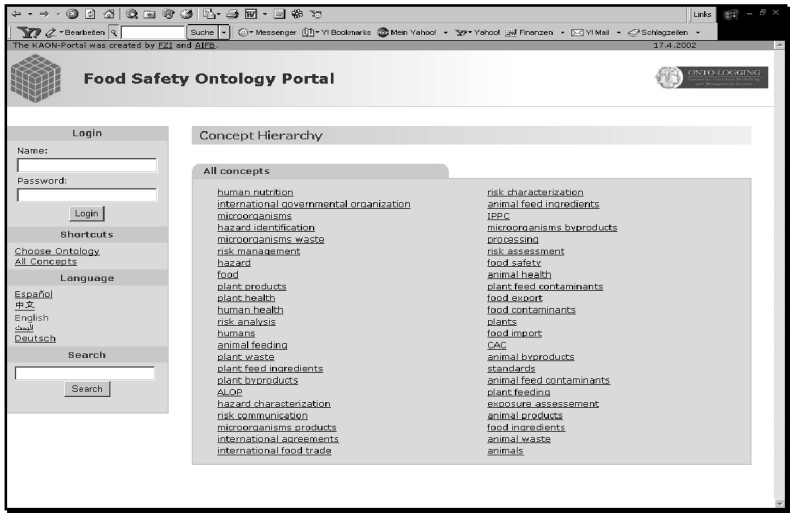



**Figure 6. Screenshot of multilingual, web based ontology browser**

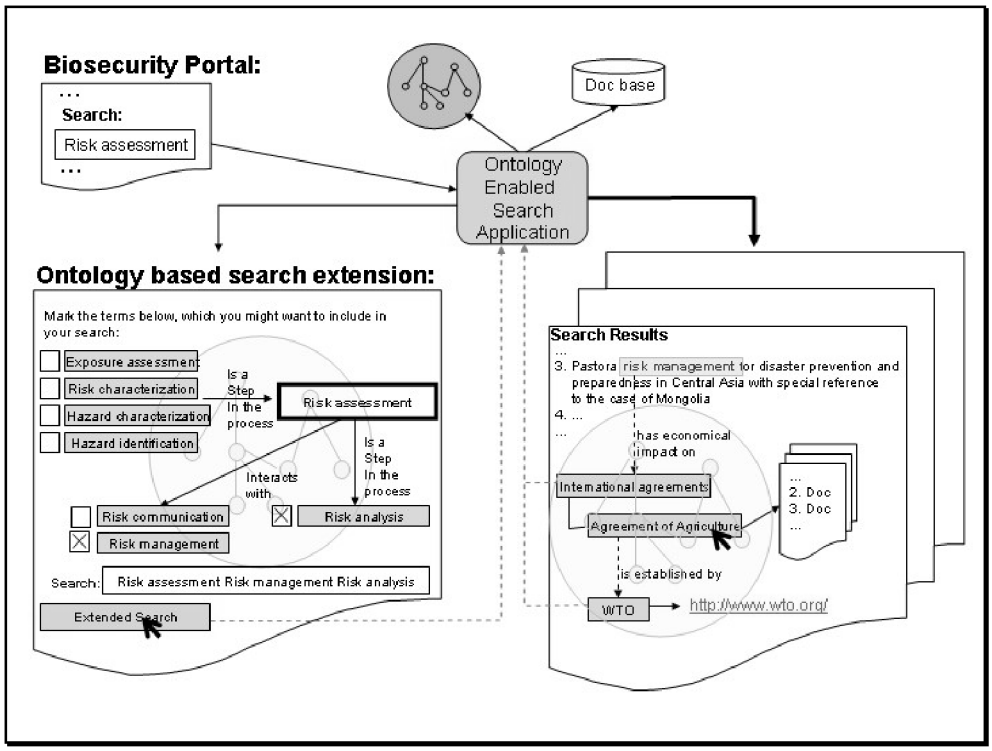



**Figure 7. Ontology based search extension and semantically structured result display**

semantic context of the search, and returns the results back to the user, giving her the possibility to extend or specify the query. The interlinked grey boxes show the conceptual neighbourhood of the search term in the Biosecurity Ontology prototype. The extended query is passed back to the application, which now searches the document base. Once again, the semantic context within the ontology can be used in order to provide the user with related results which might be of interest. The picture on the right shows an excerpt of retrieved search results. The user is provided with additional links or documents, which are related to neighbouring concepts of the initial search term. This shows how domain ontologies can be useful in knowledge discovery and providing domain relevant, semantic links among search results.

These features have yet to be implemented and evaluated in future project work. Hence, usability has not been proven at this stage.

A commercially available tool providing similar functionality (like automatic keyword search extensions and structured, enhanced result representation) is the Semantic Miner from Ontoprise[10].

In the above discussed solution, the annotation of the documents does not change and the same document bases are accessed. A further step would be, to actually annotate the documents of the domain of interest with the semantic information of the ontology. With semantic annotation, not only support in search term compilation and semantic structuring of search results can be given, but documents and their annotated content can now be interlinked semantically to provide enhanced knowledge discovery. Refer to (Staab et al. 2000b) for a detailed discussion of semantic annotation.

### 6.2. Semiautomatic, ontology based text classification

Text classification is a time-consuming task, which is typically performed manually. However, the vast amount of information on the internet makes it impossible to continue using this approach for arbitrary web documents in the future. Statistical classifiers exist and have shown quite good results using standard texts, which all follow certain patterns. A good overview about methods and evaluations is given in (Aas 1999). However, none of the methods can so far replace human classifiers, since they all lack the specialist's semantic knowledge of the domain in which the document has to be classified. Little research has been done in integrating ontological background knowledge into classical text classification methods. One attempt[11] used the freely available dictionary WORDNET[12] to serve as background knowledge for text classification with support vector

[10] http://www.ontoprise.de/com/download/semminer_iswc_submission.pdf.

[11] A research study done at the University of Karlsruhe in 2002; refer to (Pache 2002) for details.

[12] http://www.cogsci.princeton.edu/~wn/.

machines. The classifier used the News20-document-set for evaluation purposes and showed good performance. This work can now be expanded, and WORDNET can be replaced with a domain ontology, such as the Biosecurity Ontology, to evaluate the classifier against arbitrary web documents. An automatic indexing approach like this could then be used in combination with Dublin Core elements to index web pages for Semantic Web purposes.

## 7. Summary

We have presented a new approach towards domain ontology creation. The introduced framework provides a generic, reusable methodology, which can be reapplied to create domain ontologies in various fields of interest. The prototype project which has been presented in this paper showed the applicability of the methods in the biosecurity domain. We introduced a generic layered ontology modelling approach that can be used to describe any possible domain of interest. Multilingual aspects have been addressed to solve the problems of portability, usage and representation of semantic knowledge in different languages. The overall framework, we described in Section 4 and 5, provides a comprehensive methodology for domain ontology creation and is not bound to any domain specific input. Used thesauri, document sets and core ontologies can easily be replaced by equivalents from other domains. Moreover, as the open source applications are all Java-based, the used toolset providing the semiautomatic support is extremely adaptable to different needs. Obviously, the whole approach is completely portable and reusable in other domains.

We concluded our presentation, giving an outlook on further work to be done in the field of domain ontology usage. Example scenarios and applications have been addressed, giving an outlook on possible implementations of the Semantic Web: The initial motivation for building ontologies.

Acknowledgements: This project has been done in close collaboration with the AIFB[13] institute of the University of Karlsruhe (TH) in Germany. All tool supported steps have been carried out, using the freely available KAON environment, developed at this institute. We would like to express our gratitude to the KAON group (KAON) for their technical expertise in this subject. We particularly thank Raphael Volz for his sound direction, technical guidance and supervision throughout the project. We also gratefully recognize the programming support of Boris Motik, which facilitated the adaptation of the tool set.

[13] Institut für Angewandte Informatik und Formale Beschreibungsverfahren, Universität Karlsruhe (TH), Karlsruhe, Germany.

## References

Aas K., Eikvil L., Text categorisation: A survey. Technical report, Norwegian Computing Center, June 1999.

AOSProposal 2002. http://www.fao.org/agris/aos/Documents/AOS_Draftproposal.htm. June 2002.

Fernandez M., Blazquez M., Garcia-Pinar J.M., Gomez-Perez A., 1998. Building Ontologies at the Knowledge Level using the Ontology Design Environment.

Fernandez M., Gomez-Perez A., Pazos Sierra A., Pazos Sierra J., 1999. Building a Chemical Ontology Using METHONTOLOGY and the Ontology Design Environment. IEEE Expert (Intelligent Systems and Their Applications), 14(1): 37-46.

Gangemi et al., A Formal Ontological Framework for Semantic Interoperability in the Fishery Domain. February 2002.

Guarino N., Formal Ontology and Information Systems. In: N. Guarino, editor, Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98, Trento, Italy, pages 3-15. IOS Press, June 1998.

http://www.ladseb.pd.cnr.it/infor/ontology/methodology.html. June 2002.

Bozsak E., Ehrig M., Handschuh S., Hotho A., Mädche A., Motik B., Oberle D., Schmitz C., Staab S., Stojanovic L., Stojanovic N., Studer R., Stumme G., Sure Y., Tane J., Volz R., Zacharias V., KAON – Towards a large scale Semantic Web. In: Proceedings of EC-Web 2002. Aix-en-Provence, France, September 2-6, 2002. LNCS, Springer, 2002.

Kietz J.-U., Volz R., Maedche A., Extracting a Domain-Specific Ontology from a Corporate Intranet. Proc of the 2nd Learning Language in Logic (LLL) Workshop, Lissabon. September 2000.

Pache G., 2002. Textklassifikation mit Support-Vektor-Maschinen unter Zuhilfenahme von Hintergrundwissen. Studienarbeit. AIFB Universität Karlsruhe(TH), Karlsruhe, Germany. April 2002.

Palmer S., 2001. The Semantic Web: An Introduction. http://infomesh.net/2001/swintro.

RDFSchema 2002.
http://www.w3.org/TR/rdf-schema/#intro. June 2002.

Staab S., Erdmann M., Mdche A., Decker S., An Extensible Approach for Modeling Ontologies in RDF(S). In: Proceedings of ECDL 2000 Workshop on the Semantic Web, 11-22, 2000.

Staab S., Mädche A., Handschuh S., An Annotation Framework for the Semantic Web. In: S. Ishizaki (ed.), Proc. of The First International Workshop on MultiMedia Annotation. January, 30-31, 2001. Tokyo, Japan.

Volz R., Akquisition von Ontologien mit Text-Mining-Verfahren. Technical Report 27, Rentenanstalt/Swiss Life, CC/ITRD, CH-8022 Zürich, Switzerland, ISSN 1424-4691.

Welty C., Guarino N., 2001. Supporting Ontological Analysis of Taxonomic Relationships. Data and Knowledge Engineering 39(1), pp. 51-74.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 125-132*


# The MEG Registry and SCART: Complementary Tools for Creation, Discovery and Re-use of Metadata Schemas

Rachel Heery, Pete Johnston
UKOLN, University of Bath
r.heery@ukoln.ac.uk, p.johnston@ukoln.ac.uk

Dave Beckett, Damian Steer
ILRT, University of Bristol
dave.beckett@ilrt.bris.ac.uk, pldms@mac.com

## Abstract

*SCART is an RDF schema creation tool designed for use by implementers working within digital library and learning environments. This schema creation and registration tool is being developed to work in conjunction with registry software. SCART will provide implementers with a simple tool to declare their schemas, including local usage and adaptations, in a machine understandable way based on the RDF Schema specification. This tool is optimised for use by the Metadata for Education Group, projects and services within the UK providing resource discovery in the domain of education. By providing a complementary creation tool and registry the aim is to facilitate easy discovery of existing schemas already registered in a schemas registry, and to enable implementers to re-use these existing schemas where appropriate.*

## 1. Introduction

The Metadata for Education Group (MEG) Registry and Schema Creation tools aim to provide implementers of educational systems with the means to share information about their metadata schemas and to re-use existing schemas. MEG is a loose confederation of educational organisations concerned with the description and provision of educational resources at all educational levels across the United Kingdom [1]. Currently there are over sixty members of the MEG group with approximately twenty known to be active in creating schemas to describe educational resources. The existence of such a focused group offers great potential for sharing and collaboration regarding design and re-use of schemas. Facilitating 'declaring and sharing' schemas in use by members of the group will benefit their system designers, and their funders, by saving the time and effort currently spent in researching existing schemas and in re-inventing schemas.

The MEG registry work builds on previous activity within the DESIRE and SCHEMAS projects [2, 3] which established data models for declaring schemas and local usage within a schema registry, and implemented prototype registries [4, 5, 6]. The MEG registry development is based on the DESIRE data model, but is a completely new implementation which seeks to address some of the problems of sustainability and scalability encountered with the DESIRE approach. The DESIRE registry was implemented as a relational (MySQL) database. The aim within the MEG development is to explore the benefits, and drawbacks, of implementation using specifications and tools emerging from the Resource Discovery Framework (RDF). The status of the development is work in progress, with a final delivery date planned for September 2002. Prototypes are now available which can be accessed and used for demonstration, with draft schemas loaded into the Registry [7].

A schema creation and registration tool (SCART) is being developed to work in conjunction with the new MEG Registry software. The SCART will enable implementers to declare their schemas, including local usage and adaptations, in a machine understandable way. This tool is optimised for use by projects and services providing resource discovery in the domain of education, whether for discovery of information or learning objects. By providing a complementary creation tool and registry the aim is to facilitate easy discovery of existing schemas already registered in a schemas registry, and to enable implementers to re-use these existing schemas where appropriate, whilst creating new usages where required. The SCART is designed to interact with the

MEG Registry, so both are being developed in parallel.

The MEG registry will provide browsing and searching facilities for all data elements contained within the registered schemas, and schemas can be entered or updated by use of the SCART. Experience from previous prototypes (DESIRE, SCHEMAS) has established that it is vital for implementers to have this facility for entry and update under their local control: it is not scalable in terms of effort to manually enter schemas centrally. This imposes the requirement that local implementers construct their schemas in a well-formed manner, and the SCART is designed with this in mind. An additional benefit is that once a well-formed schema has been produced locally it can be used for local applications, and provided to (or gathered by) other registry-like applications. That is, the SCART and the MEG registry are loosely coupled, in that schemas created using the SCART may be used by other applications and schemas may be prepared using other tools and submitted to the registry.

In order to follow existing recognised standards as far as possible, and to take advantage of emerging open source tools the MEG developments have been based on the RDF Schema specification [8, 9]. RDF provides a common data model which is particularly appropriate for exchange of data with unknown semantics by enabling common naming and identification of data.

This paper will provide some context to the MEG developments, then go on to describe the data model used within the SCART and Registry. Lastly some brief detail will be given of the design and features provided within the SCART, while keeping in mind that this is at present in prototype, with delivery scheduled for September 2002.

## 2. Background

A short overview of activity within the Semantic Web regarding vocabulary sharing will give some context to the MEG registry work, as this has informed development of the SCART and MEG registry. The vision of the Semantic Web is built on an infrastructure of interoperable metadata, where software can infer semantic meaning to 'unknown' metadata, albeit with an acceptance that the understanding of these semantics may be partial. The Semantic Web envisages software being able to treat the Web as 'a global database' [10] where data (and metadata) can be fetched and manipulated. In order to achieve this, there needs to be a way for metadata to be exchanged both at syntactical and semantic levels.

The Resource Description Framework (RDF) provides a common data model for making statements about resources, and a means of expressing those statements in a common syntax (XML). This combination of model and syntax means that independently created statements describing the same resource can be shared and 'merged'. Such data aggregation offers a powerful means to re-use existing (RDF/XML) metadata that resides on the Web or in accessible RDF compliant databases. RDF provides a data model. However there still needs to be consensus on identification and naming to enable shared use of metadata, both for the naming and meaning of 'properties and classes' (data elements), and for identification of the resources being described. The importance of naming has been acknowledged by the TAP project [11] and it is a complex area which will take time to solve. However we see deployment of schema creation tools and schema registries as a means to assist in reaching a common approach to naming data elements, which can then be shared by re-use.

Within the digital library and wider Semantic Web community there has been some exploration of options for declaring and sharing metadata schemas. Various approaches to establishing common vocabularies have been put forward:

**Relying on dominant market forces**, whereby the core vocabulary of e-commerce will prevail [11].

**Mapping between core data elements** whether mapping of data elements [12] or more complex ontology mapping [13].

**Enabling 'base-line' interoperability** by agreement on a minimal metadata element set, whereby heterogeneous metadata is normalised to a minimum common data element set [14].

These various options may be seen as transitory stages to a more sophisticated solution, or the solution may be a more complex amalgamation of all approaches. However it is clear that all of these approaches require flexibility and extensibility, which in turn will require manipulation and mapping of different vocabularies. Registries are seen as providing such services. Within the digital library community a first step might be to enable declaration of vocabularies in use, in particular to publish and share vocabularies. For some time metadata registries have been seen as a way to do this, to encourage re-use where appropriate:

> *… registries will need to be managed, coordinated, and ultimately connected. Registries will define the elements of metadata schemas in a machine-readable syntax (e.g., RDF) and offer authoritative listings of legal values, local extensions, mappings to other schemas, and guidelines for good usage. They will serve both humans, with readable text, and programs, with structured content that can automatically be parsed. Their role will be both to promote and to inform, thereby encouraging the use of standard formats. [15]*

Acknowledging the need to reach consensus on data elements that can be used in common, it seems likely that this might best be achieved incrementally by agreeing common vocabularies amongst shared 'communities of interest'.

*The standard vocabulary need not all come from one source. We provide a core kernel of terms, which can be extended by communities and their applications. Some of the terms defined by the applications will over a period of time get absorbed into the kernel. The use of XML namespace qualifications allows data providers to define their own extensions. The evolution of this ecology of names will likely mirror the evolution of operating system*

*APIs, wherein, over a period of time, the operating systems incorporated the APIs offered by the more successful platform-like applications running on the operating systems [11].*

## 3. The role of the Registry

The functionality of particular metadata registry implementations differ, but the overall role can be encapsulated as facilitating extensibility and interoperability in the context of networked services. Within the digital library community there have been a variety of approaches, so for example some registries have been human readable only, whereas others machine-readable. Some registries offer descriptive information of element sets whereas others provide search and browse access to data elements. In some cases, those elements are from schemas owned by a single 'registration agent'; in other cases, they are drawn from many schemas from various sources. Some registries might allow the interpretation of different metadata element sets by means of crosswalks, mappings or translations. Registries might provide a service to a specific community of practice, such as the education domain, or the museum sector, or might be focused on a group of implementations with a common business model. A brief overview of registry activity is included in a recent account of the DCMI Registry activity [16]. The DCMI Registry is an example of a standards making body providing information about its element sets by means of a registry service [17].

The MEG Registry is a formal system that discloses authoritative information about the semantics and structure of the data elements within the registered schemas. The MEG registry uses schemas primarily to provide a descriptive or documentary function, and it should be noted that the validation of instance data against schemas is not part of that function. The MEG Registry will provide information about its contents to both humans and software. This means the information within the registry needs to be stored in a syntax that is machine-readable as well as in human readable form. Users of the Registry System would typically be implementers seeking appropriate schemas, developers comparing schemas, publishers of standards, and metadata creators seeking assistance in using particular schemas correctly.

Usage scenarios for the MEG Registry System include:

**Publishing a description of an Element Set:** A UK organisation provides a resource discovery service for Web-based educational materials that utilises a simple metadata schema developed specifically for that purpose. The organisation wishes to publish this information to the MEG community via the registry. Using the SCART the Element set publisher can create an Element Set description, and add descriptions for each Element. Encoding Schemes can be added to Elements. On completion the Element Set can be saved locally and submitted to the Registry

**Publishing a description of an Application Profile** A UK organisation provides a resource discovery service for Web-based educational materials. That service utilises a simple metadata schema that uses a number of Elements drawn from the cross-domain Element Sets of the Dublin Core Metadata Initiative; a domain-specific Element which was created by another portal service for their own schema; and a number of new Elements specific to this service. The organisation has developed a number of controlled vocabularies for several of the Elements in this schema; the service also specifies the use of some standardised forms for dates and identifiers within metadata instances. The organisation wishes to publish this information to the MEG community via the registry. In the terms of the registry data model, this organisation's schema is an Application Profile.

**Indexing a standard schema for an Element** Set An international standards body makes schema for their cross-domain Element Set available in RDF/XML on their Web server. Various MEG members wish to 'use' Elements from the Element Set in their Application Profiles. Either the representative of standards body or the registry administrator can use SCART to add the schema to the Registry.

**Exploring Element Usage** A schema developer wishes to survey the usage of the DCMI 'audience' element, and particularly the use of any controlled vocabularies to control values of this element.

## 4. The MEG Registry data model

Underpinning the data model for the MEG registry is the recognition that implementers deploy and adapt 'standard' metadata Element Sets in a pragmatic way. While 'standard' schemas are widely available, use-oriented adaptations, which are often localised and service-specific, tend to be less visible. Researchers on schema usage have introduced the idea of the 'application profile' as a means of capturing this information on adaptations and constraints of Element Set usage [18].

The data model for the MEG registry is designed to support the description of the following classes of entity and the relationships between instances of those classes. Descriptions of all instances include a unique identifier (or token) - for use (primarily at

least) by software tools - and a label or title for the human reader.

**Elements:** the formally defined terms which are used to describe attributes of a resource. The description of an

Element must include a unique identifier, a name or label, and a description of its meaning. It may include information on the usage of the Element and information on the relationship between this Element and semantically-related Elements.

**Element Sets:** sets of functionally-related Elements which are defined and managed as a unit. The description of an Element Set must include a unique identifier, a title, a textual description of its intended scope/area of use, and the name (URI) of an XML Namespace associated with the Element Set. It may include version information, a classification of the Element Set and references to descriptions of the Element Set.

**Usages of Elements:** in the context of particular applications. The description of an Element Usage must include a unique identifier and the identifier of an Element. It may include:

- a new name or label for the Element in this application context;
- a description of the meaning of the Element in this context (the Element Usage may refine the definition of an Element to make it narrower or more specific to an application context);
- a description of the obligation to use the Element, and/or any constraints on its occurrence, in this application context;
- a description of constraints on the value of the Element in this application context, either as data type specifications or more narrowly through association with one or more Encoding Schemes (see below).

**Application Profiles:** sets of functionally-related Element Usages, created for the purpose of a particular function or application and managed as a unit. An Application Profile may include Element Usages of Elements from one or more Element Sets. The description of an Application Profile must include a unique identifier, a title and a textual description of its intended scope/area of use. It may include version information, a classification of the Application Profile, references to external descriptions of the Application Profile, and references to XML Schema based on the Application Profile.

**Encoding Schemes:** mechanisms that constrain the value space of Elements. The description of an Encoding Scheme must include a unique identifier, a name or label and a textual description of its intended use. It may include version information, a classification of the Encoding Scheme, and references to external descriptions of the Encoding Scheme. Encoding Schemes may be of two types:

- a Scheme which enumerates a list of permitted Values: the list of Values may be recorded by the registry (see below);
- a Scheme which specifies a set of rules that define or describe permitted values: such rules cannot be captured by the registry, but can be indicated by a reference to an external description of the Encoding Scheme.

**Values:** the individual Values which an Encoding Scheme enumerates may be recorded. The description of a Value must include a unique identifier and a label. It may include a textual description providing more information about the Value. The practicality of recording Values within the MEG Registry may depend on the size of the "vocabulary" and whether or not it already exists in a suitable machine-processable form.

**Agencies:** persons or organisations responsible for the ownership or management of Element Sets, Application Profiles and Encoding Schemes. The description of an Agency must include a unique identifier and a name; it may include a reference to an external source of further information.

The principal relationships between entities are represented graphically in Figure 1, with an indication of whether the relationship is many (m) to one (1).

The main points to note on the relationships between entities in this diagram are:

- Each Element Set, Application Profile and Encoding Scheme must be associated with exactly one Agency responsible for its maintenance; an Agency may be responsible for multiple Element Sets, Application Profiles and Encoding Schemes;
- An Element Set contains multiple Elements; and an Element must be a member of exactly one Element Set;
- An Element may be associated with multiple Encoding Schemes which specify constraints on its value; and an Encoding Scheme may be associated with multiple Elements;

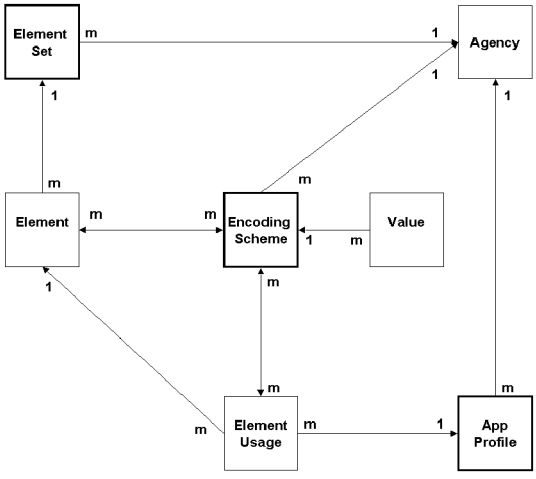



**Figure 1. Registry data model**

- An Element may be described as a semantic refinement of a second Element; and an Element may have multiple refinements;
- An Application Profile contains multiple Element Usages; and an Element Usage must be a member of exactly one Application Profile;
- An Element Usage must use exactly one Element; but an Element may be the object of multiple Element Usages;
- An Element Usage may specify the use of multiple Encoding Schemes; and an Encoding Scheme may be deployed in multiple Element Usages.

This model is based on that used by the DESIRE metadata schema registry [4]. The MEG registry also builds on the experience of the SCHEMAS project which suggested conventions for describing Application Profiles in machine-processable form using the RDF model [19, 20]. In particular, the MEG registry adopts the suggestion by the SCHEMAS project that the entities described here as Element Usages might usefully be modelled as resources, and the RDF vocabulary used by the MEG registry defines a class "reg:ElementUsage" for this purpose. Elements are modelled as resources of type rdf:Property.

Figure 2 is a graphical representation of how an "Element" and an "Element Usage" might be described using this data model.

The lower part of this diagram represents the description of the Element with the identifier http://purl.org/dc/elements/1.1/title, which is part of an Element Set defined by the Dublin Core Metadata Initiative, i.e. the element often referred to by the XML qualified name "dc:title". DCMI assigns this Element a name or label, the string "Title", and provide a definition of the Element as the string " A name given to the resource". This is represented in the diagram by labelled arcs linking a node representing the Element to two separate nodes representing these strings. A fourth linked node makes explicit the "type" of this resource. So the lower part of the diagram represents the statements:

- The resource http://purl.org/dc/elements/1.1/title is of type rdf:Property;
- The resource http://purl.org/dc/elements/1.1/title has a label, "Title";
- The resource http://purl.org/dc/elements/1.1/title has a definition, "A name given to the resource";

The upper part of the diagram represents the description of a Usage of this same Element in the context of a particular application. An implementer has chosen to adopt the "dc:title" Element but to modify the human-readable label of the Element and also to refine the semantics of the Element to make them more specific to the context of the application.

In the MEG registry model, the Element Usage is represented as a second resource - a separate node in the diagram. The relationship between the Usage and the Element is represented by the arc between the two nodes, and that arc is labelled to identify the nature of that relationship. The additional arcs and nodes represent the application-specific label and definition which the Element Usage prescribes for the Element http://purl.org/dc/elements/1.1/title. And a final linked node makes explicit the type of the Element Usage resource. So there are now four additional statements:

- The resource http://example.org/elementUsage/title is of type reg:ElementUsage;
- The resource http://example.org/elementUsage/title "uses" the resource http://purl.org/dc/elements/1.1/title;
- The resource http://example.org/elementUsage/title has a label, "Name";
- The resource http://example.org/elementUsage/title has a definition, "The name of the location".

A more complex example of an Element Usage might introduce constraints on the value of the Element in the context of the application by mandating the use of specified Encoding Schemes. These would be represented by additional nodes linked to the Element Usage node, i.e. additional statements about the Element Usage, of the form:

- The resource http://example.org/elementUsage/date specifies use of the resource http://purl.org/dc/terms/W3CDTF (which is of type reg:EncodingScheme).

Further, to simplify the diagram above, some of the relationships that would be mandatory within the context of the registry - and would be enforced by the schema creation tool - are not illustrated here. e.g. the Element node would have a further arc to a node representing a resource of type "Element Set" and the Element Usage node would have a further arc to a node representing a resource of type "Application Profile".

Some points to note include firstly, that the RDF vocabulary for the MEG registry defines a property "reg:uses" to express the relationship between an Element Usage and an Element. The data model

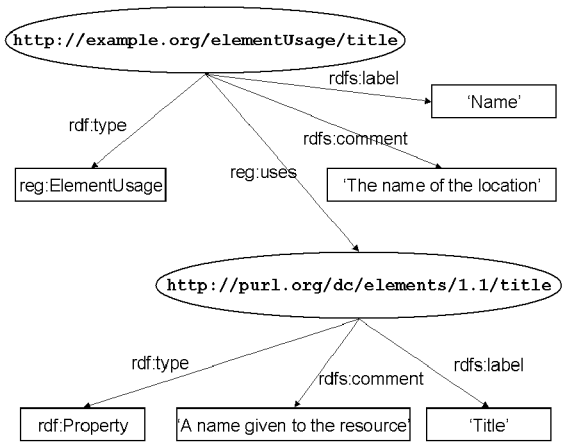



**Figure 2. Using an Element**

specifies that the value of a "reg:uses" property must be a resource of type Element. The value of a "reg:uses" property cannot be a resource of type Element Usage. The schema creation tool enforces this constraint. Element Usages are, however, assigned URIs and this does open up the possibility that, in a distributed environment, the creators of new Application Profiles might make their own statements about Element Usages previously created by others. The SCHEMAS project noted that such a possibility risks "semantic drift", but also that it may be difficult to avoid in the decentralised context of the Web [20].

Secondly, the "Element Usage" class defined in the MEG registry vocabulary is not defined as a sub-class of the "rdf:Property" class i.e. Element Usages are not RDF properties and resources of this type cannot be used as predicates in RDF statements. So, where an Application Profile specifies an Element Usage of the Element http://purl.org/dc/elements/1.1/title, the creator of instance metadata conforming to this Application Profile continues to use the term http://purl.org/dc/elements/1.1/title as the predicate in their RDF statements.

Thirdly, an Element Usage can specify a narrowing of the definition of an Element in a "standard" Element Set; an implementer might achieve a similar result by defining a new Element (in a "local" Element Set) and specifying that this Element is an "element refinement" of a standard Element. It may be useful to explore further the advantages and disadvantages of the two approaches.

## 5. SCART software development

The main capabilities required by the client software (SCART) were to:

Create and edit application profiles, element sets and encoding schemes.

Encourage re-use of existing elements and encoding schemes.

Allow submissions to a remote registry server.

The client and remote registry were required to use a common data model based on RDF Schema specification. The creation of RDF data is clearly impractical for non-experts, so the client needed to simplify the process greatly. Happily the data model is relatively simple. Encouraging reuse was the second major issue. The natural solution is to ensure that it is easier to find suitable pre-existing items than it is to create new ones. The MEG Registry will, when in production, conveniently provide a store of existing elements and encoding schemes. In the future it is expected other compatible schema registries will become available too. So SCART provides the user with the facility to search remote registries, and results are presented so that users can make informed decisions as to whether found items are appropriate for their application profiles.

Existing software offerings were considered as a basis for the client. Currently there are few options for RDF authoring and three applications were investigated: RDFAuthor [21], IsaViz [22], and Protégé [23]. RDFAuthor and IsaViz are similar applications, presenting a graphical representation of the RDF data model, that is a graph structure with nodes. For the purposes of SCART neither application was thought suitable. Both probably could have been augmented to talk to registries, however although they hide the syntax of RDF from the user neither hides the graph model. Using either tool would require a familiarity with the registry data model that is unreasonable to expect in the target audience.

Protégé is quite different in scope and intention. It is a complex tool allowing users to create ontologies, use these ontologies to create interfaces for entering instance data, and query that data. Protégé includes plugins for the DAML [24] and RDF Schema vocabularies to describe the ontology. The latter was particularly interesting since the registry uses RDF Schema. Protégé showed promise for schema creation; and submission to a remote registry probably could have been accomplished with an add-on tool; however re-use of existing data elements would have been problematic. In addition Protégé's notion of an ontology is a great deal stronger than required by the registry and the redundant functionality would result in a confusing interface for the target audience. In summary, Protégé, IsaViz, and RDFAuthor essentially are general purpose tools, and whilst suitable for the tasks for which they are designed, the MEG project needed a tool tailored for its intended audience.

SCART therefore was custom built for the project, using Java, which had a clear advantage due to its multi-platform nature. SCART is known to run on Windows 2000, Mac OS X and Linux. The promise of Java is that it will run on other platforms (e.g. other Windows versions, BSD). It should be noted that the user interface toolkit in Java (sometimes called 'Swing') has revealed some quirks, but generally proved useful, as did the RDF toolkit for Java, Jena [25]. A walk through of the functionality of the SCART is available on the MEG Web pages showing in detail the process for creating application profiles, schemas and encoding schemes [7]. Here only a brief mention of some features is possible.

SCART supports multiple documents, each a self contained window associated with a Registration Agency. When a new document is created the user has to supply a name for the agency, and an identifier (a URI). The identifier is particularly important due to the nature of RDF. However this is the only time the user is required give an object an identifier as a default identifier will be created automatically if left unspecified by the schema creator.

A 'Search Registry' window provides the re-use incentive to a schema creator. It provides an interface to 'external' elements and encoding schemes whether these are located at a remote registry, or in a local file

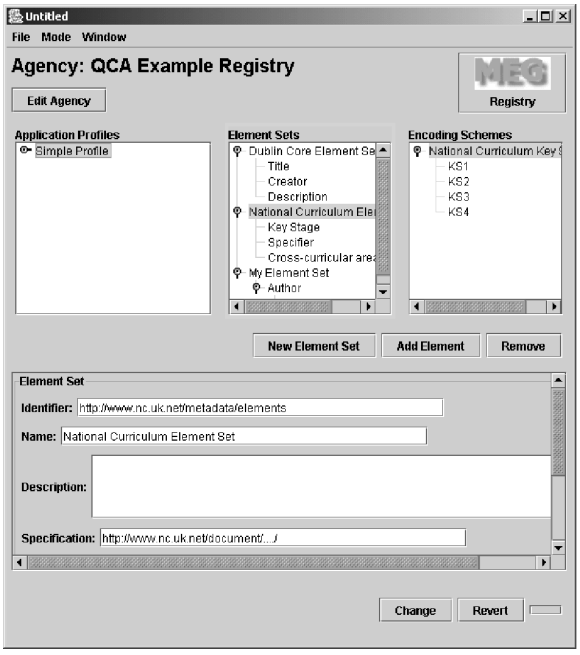



**Figure 3. Creating an application profile with the SCART**

allowing offline use. Searching is carried out using the HTTP protocol and, of course, other registries searched need to be structured according to a common data model. Users can perform simple keyword searches to find elements and/or encoding schemes. Results are displayed so they can be dragged and dropped into the application profile window.

Drag and drop is used extensively throughout SCART. Encoding schemes and elements can be dragged to any place where their inclusion makes sense. In this way the user can create application profiles made up of existing data elements.

Separate windows are provided for the main tasks. One for creating application profiles and element usages, one for element sets, and one for encoding schemes. The most typical use of the tool within the MEG context will be for creating application profiles so the default view in the window is of the 'application profile', however when required two further views to the document window are made available for creating elements sets and encoding schemes.

Resulting schema can be saved locally (in RDF/XML), then submitted to the registry, or reloaded to the SCART later for further editing. On completion the schema can also be made available for harvesting by other applications, with the advantage that other agencies can reuse the data.

## 6. Conclusions

The primary purpose for describing Application Profiles and Element Usages is to provide a means by which schema implementers can disclose information about service-specific or application-specific practice in using metadata schemas. The information serves primarily a 'documentary' function [26, 20]. The ability to express this information in the form of machine-processable schemas facilitates the exchange and reuse of that information. This suggests that it is possible for metadata schema registries to broaden their scope to index and publish not only the descriptions of Element Sets provided by 'standards bodies', but also information on the local, 'real world' experience of implementing those Element Sets.

Information on implementation forms a larger and more rapidly changing set of data than the descriptions of the relatively static standard Element Sets. The effective capture of this information by services such as schema registries will depend on the provision of tools appropriate to the distributed and decentralised nature of the environment. The development of the schema creation tool seeks to address that challenge.

Information on implementation is also potentially much less uniform than the description of standard Element Sets. The data model outlined here has emerged primarily from the experience of working with the Dublin Core Metadata Element Set, which is a small, simple Element Set. We believe that the principles on which the work is based are extensible to other more complex schemas, but that hypothesis remains to be tested. The provision of a registry for the MEG community, with a number of implementers of schemas based on the IEEE Learning Object Metadata Specification and the IMS Learning Resource Meta-data Specification, will provide an opportunity to do so.

## Acknowledgements

This work was made possible through funding from the JISC and BECTa The authors would like to acknowledge Tom Baker and Manjula Patel for their contribution to discussions of registries and schema creation tools within the context of previous projects.

## References

[1] MEG, *Metadata for Education Group*, 2002. Available from: <http://www.ukoln.ac.uk/metadata/education/> (Last visited: June 2002).

[2] DESIRE, *Desire Project*. 2000. Available from: <http://www.desire.org/> (Last visited June 2002).

[3] SCHEMAS Forum, *The SCHEMAS Forum*, 2001. Available from: <http://www.schemas-forum.org/> (Last visited: June 2002).

[4] Heery, R., Gardner, T., Day, M., & Patel, M.,

*DESIRE metadata registry framework,* 2000. Available from <http://www.desire.org/html/research/deliverables/D3.5/> (Last visited June 2002).

[5] DESIRE, *Desire Metadata Schema Registry*, 2000. Available from: <http://desire.ukoln.ac.uk/registry/> (Last visited June 2002).

[6] SCHEMAS Forum, *The SCHEMAS Forum Registry,* 2001. Available from: <http://www.schemas-forum.org/registry/> (Last visited: June 2002).

[7] MEG, *MEG registry project*, 2002. Available from: <http://www.ukoln.ac.uk/metadata/education/regproj/> (Last visited: June 2002).

[8] Brickley, D. and Guha, R.V. Editors, *Resource Description Framework (RDF) Schema Specification 1.0. World Wide Web Consortium W3C Candidate Recommendation, 27 March 2000,* W3C, 2000. Available from <http://www.w3.org/TR/2000/CR-rdf-schema-20000327/> (Last visited June 2002).

[9] Brickley, D. and Guha, R.V. Editors, *RDF Vocabulary Description Language 1.0: RDF Schema. World Wide Web Consortium W3C Working Draft, 30 April 2002.* Work in progress, W3C, 2002. Available from <http://www.w3.org/TR/2002/WD-rdf-schema-20020430/> (Last visited June 2002).

[10] Berners-Lee, T., *Semantic Web Road map*, 1998. Available from: <http://www.w3.org/DesignIssues/Semantic.html> (Last visited 22 June 2002).

[11] Guha, R.V. and McCool, R., *A System for integrating Web Services into a Global Knowledge Base,* 2002. Available from <http://tap.stanford.edu/sw002.html> (Last visited June 2002).

[12] Blanchi, C. and Petrone, J., Distributed Interoperable Metadata Registry. *D-Lib Magazine,* Volume 7 Number 12 (Dec 2001). Available from: <http://www.dlib.org/dlib/december01/blanchi/12blanchi.html> (Last visited: June 2002).

[13] Lagoze, C. and Hunter, J., The ABC Ontology and Model. *Journal of Digital Information*, Volume 2 Issue 2 (Nov 2001). Available from: <http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/> (Last visited: June 2002).

[14] Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. Editors, *Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0 of 2002-06-14*. OAI, 2002. Available from: <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm> (Last visited: June 2002).

[15] EU-NSF Working Group on Metadata, *Metadata for digital libraries: a research agenda*. Draft 10. Le Chesnay: ERCIM, 1998. Available from: <http://www.ercim.org/publication/ws-proceedings/EU-NSF/metadata.html> (Last visited: June 2002)

[16] Heery, R. and Wagner, H., A Metadata Registry for the Semantic Web. *D-Lib Magazine,* Volume 8 Number 5 (May 2002). Available from: <http://www.dlib.org/dlib/may02/wagner/05wagner.html> (Last visited: June 2002).

[17] DCMI Registry Working Group, *The Dublin Core Metadata Registry*. Prototype, 2002. Available from: <http://wip.dublincore.org:8080/dcregistry/index.html> (Last visited: June 2002).

[18] Heery, R. and Patel, M., Application Profiles: mixing and matching metadata schemas. *Ariadne* 25 (Sep 2000). Available from <http://www.ariadne.ac.uk/issue25/app-profiles/> (Last visited June 2002).

[19] SCHEMAS Forum, *The SCHEMAS Forum Registry: sample RDF encodings of Application Profiles*, 2002. Available from: <http://www.schemas-forum.org/registry/schemas/> (Last visited: June 2002).

[20] Baker, T., Dekkers, M., Heery, R., Patel, M. and Salokhe, G., What terms does your metadata use? Application profiles as machine-understandable narratives. *Journal of Digital Information* Volume 2 Issue 2 (Nov 2001). Available from: <http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Baker/> (Last visited: June 2002).

[21] Steer, D., *RDF Author,*. 2002. Available from: <http://rdfweb.org/people/damian/RDFAuthor/> (Last visited: June 2002).

[22] Pitriga, E., *IsaViz: a visual authoring tool for RDF*, 2002. Available from: <http://www.w3.org/2001/11/IsaViz/> (Last visited: June 2002).

[23] *Protégé*, 2002. Available from: <http://protege.stanford.edu> (Last visited: June 2002).

[24] DAML, *DARPA agent markup language homepage*, 2002. Available from: <http://www.daml.org/> (Last visited: June 2002).

[25] Hewlett Packard Labs, *HPL Semantic Web activity*, 2002. Available from: <http://www.hpl.hp.com/semweb/jena-top.html> (Last visited: June 2002).

[26] SCHEMAS Forum, *The SCHEMAS Forum: A Retrospective Glossary*, 2001. Available from: <http://www.schemas-forum.org/info-services/d74.htm> (Last visited: June 2002).

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 133-138*


# Does metadata count? A Webometric investigation

Alastair G. Smith
School of Information Management
Victoria University of Wellington
New Zealand
Alastair.Smith@vuw.ac.nz

## Abstract

*This study investigates the effectiveness of metadata on websites. Specifically, the study investigated whether the extent of metadata use by a site influences the Web Impact Factor (WIF) of the site. The WIF is a Webometric measure of the recognition that a site has on the web. WIFs were calculated for two classes of sites: electronic journals and NZ University Web Sites. The most positive correlation was found between the substantive WIF of the electronic journal sites and the extent of Dublin Core metadata use. The study also indicates a higher level of metadata use than previous studies, but this may be due to the nature of the sites investigated.*
**Keywords:** *metadata, effectiveness, evaluation, Web Impact Factors, search engines, electronic journals, university web sites.*

## Introduction

There has been much discussion of the value of metadata in providing intellectual access to digital objects. In library and information management circles the value of metadata is taken as a given. However there has been relatively little empirical evaluative investigation of the benefits of metadata. Is metadata simply a "good thing" along with motherhood and apple pie, or can its value in enhancing the value of sites, and intellectual access to them, be demonstrated objectively?

We do know that on the World Wide Web relatively few sites use metadata (Lawrence & Giles 1999). When metadata is used, it is often not used effectively. For instance a metadata template may be copied across sites without being modified to reflect the intellectual content of the site. As an example the Intellectual Property Office of NZ site (http://www.iponz.govt.nz) shares metadata with motor vehicle registry, so entry page for intellectual property office includes the inappropriate keyword "motor vehicles".

How could we evaluate the impact and benefits of metadata? Two possible approaches present themselves.

We could investigate the impact of metadata on searching: carry out an empirical investigation of the effectiveness of searches for documents which have metadata attached, and compare this the retrieval of documents without metadata. Such research needs to take account of issues relating to relevance, and evaluation of search engine performance (Oppenheim et al. 2000). In particular such research would need to choose search terms independent of the language used in the target documents, and in their metadata. Such a study has been carried out (Henshaw & Valauskas 2001), in which the retrieval of articles from an electronic journal were compared before and after the addition of metadata; the results indicated that metadata in itself did not impact on the ranking or retrieval by Internet search engines.

Another approach is to evaluate the impact factor of websites and relate this to the extent of metadata use. A Web Impact Factor (WIF) is a relatively new measure of the extent to which a site is linked to by other sites, and is analogous to a citation count in the print environment. Broadly, it is a measure of the extent of the reputation of a site, the extent to which it is linked to and recognised by other sites.

WIFs are part of the methodology of webometrics. Björneborn (Björneborn 2002) defines 'webometrics' as: "The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and informetric methods".

The idea of applying bibliometric techniques to the web was proposed by Almind and Ingwersen (Almind & Ingwersen 1997). Ingwersen (Ingwersen 1998) proposed the measure of the WIF, analogous to the Journal Impact Factor in the print publishing environment. Broadly defined, a Journal Impact Factor is the ratio of citations made to a journal to the number

of citable articles in the journal. Ingwersen proposed that the WIF should be defined as the ratio of links made to a website, to the number of pages at the website. Ingwersen distinguished between:

- the simple WIF, the ratio of all links to the number of pages;
- the internal WIF, the ratio of internal links within the site to number of pages;
- the external WIF, the ratio of links made from external sites to the target site, to the number of pages at the site.

In practice the external WIF appears to be the most valid measure of impact for a site. It is noteworthy that this is similar to the Google concept of page rank (Brin & Page 1998). WIFs can be calculated from data derived from searches on web based search engines, for instance AltaVista. While most of the major search engines can in theory be used for webometric study, in practice AltaVista provides the best combination of a large database, consistent results, and boolean logic for combining complex search results. Problems with Altavista in an earlier study (Smith 1999) appear to have been overcome.

Thelwall (Thelwall 2000) has attempted to correlate WIFs with external measures of research output of British universities, and found that a WIF that concentrated on research based pages gave the highest correlation with external measures. This result was broadly confirmed in a parallel study of Australasian Universities (Smith & Thelwall 2002).

This paper describes an exploratory webometric study, attempting to establish if there is a correlation between the impact factors of electronic journals and of New Zealand University web sites; and the extent to which metadata is used on the site.

The study also tested the extent to which links made to e-journals were to the e-journal as an entity (for instance from a list of e-journals) or to specific articles or other substantive material in the e-journals (the equivalent of a print citation to a specific article).

## Methodology

A number of e-journal sites were surveyed. 33 E-Journals were selected from a range of sources, including the *Electronic Journal Miner*, http://ejournal.coalliance.org/, using the following criteria:

- full text of journal articles were freely accessible on the web;
- the journals were pure e-journals, i.e. no print equivalent that could "pollute" citations;
- the journals were refereed, with at least some scholarly research articles;
- the journals had a distinctive URL that could distinguish the content of the e-journal.

For each e-journal, the following data was gathered:

- [P] number of pages spidered by AltaVista (host:{url} or url:{url} depending on whether the URL was a domain (e.g. for firstmonday.org, the host command was used) or a subdirectory (e.g. for dlib.org/dlib, the url command was used).
- [X] number of external links made to the e-journal (link:{url} and not host:{url} or link:{url} and not url:{url}).
- Proportion of pages spidered by AltaVista that contained metadata (keyword, or description) or DC metadata. This was done by sampling the first 10 URLs in the AltaVista hit list. In advanced search mode AltaVista presents results in random order, so this is a valid sample. In retrospect a more thorough study would include more URLs, but this was felt at the time to be a sufficient sample to indicate the extent of metadata use by the site. No attempt was made to judge the quality or quantity of metadata; pages were simply counted according to whether keyword or description metadata was present, and whether it was in DC format.
- [L] Proportion of linking pages that linked to substantive content in the e-journal. Many links are made to an e-journal from lists of e-journals, which does not imply impact; references made to specific articles and other content are potentially a better indication of impact.
- A similar data gathering exercise was followed for the eight NZ University websites, except that no attempt was made to assess the substantive nature of the links.

From this data, two measures of impact factor were calculated:

- The "original" external WIF, the ratio P/X.
- The substantive WIF, the ratio of links made to substantive content in the e-journal, the ratio P/X*L/100. This measure is closer to that of a journal impact factor, since it excludes links made to an e-journal from lists, which do not imply a measure of recognition.

## Results

### Electronic Journals

The raw data from the study is provided in appendices 1 & 2. Interestingly, comparisons using the "original" external WIF, the ratio of links from external sites to the e-journal to the number of pages at the e-journal, show little evidence that extent of metadata enhances the impact factor of the journal.

| | |
|---|---|
| Average WIF for no metadata | 6.71 |
| Average WIF with metadata | 4.27 |
| Average WIF with DC metadata | 5.33 |

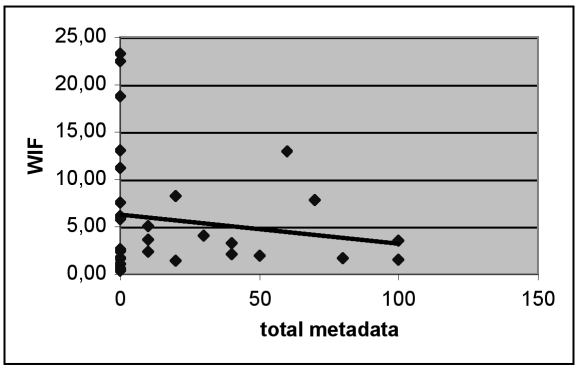



**Figure 1. "Original" WIF against total metadata**

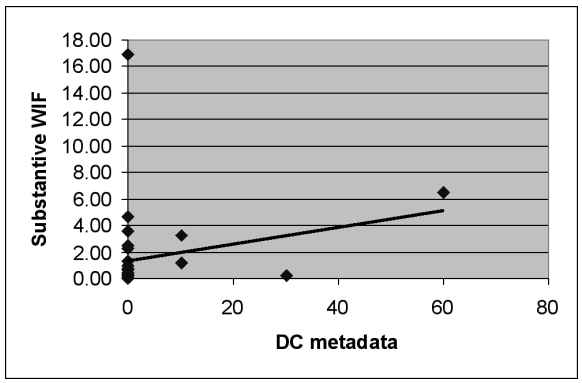



**Figure 3. Substantive WIF vs DC metadata**

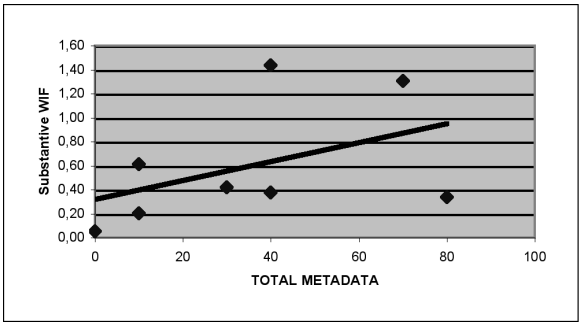



**Figure 2. Substantive WIF against total metadata**

A graph of the total metadata against the WIF indicates a slightly negative correlation (an Excel correlation coefficient of –0.15):

However the substantive WIF gave more support for the value of metadata. A graph of the total metadata against the substantive WIF indicates a slight positive correlation (an Excel correlation coefficient of 0.06), while a graph of the DC metadata against the substantive WIF indicates a stronger correlation (Excel correlation coefficient of 0.19):

| | |
|---|---|
| Average subst WIF with no metadata | 1.46 |
| Average subst WIF with metadata | 1.90 |
| Average subst WIF with DC metadata | 2.77 |

The data also allowed an estimate of the proportion of pages in e-journals that use metadata.

| | |
|---|---|
| % pages with metadata | 19.61 |
| % pages with DC metadata | 2.94 |

This contrasts with the Lawrence & Giles (1999) estimate of 0.3% of sites using DC metadata.

**NZ University websites**

A similar study was carried out of NZ University websites. No attempt was made to distinguish "substantive" links from others: this is too subjective when sites do not have clearly distinguishable information units in the way that e-journals have, and Altavista does not clearly distinguish "research pages" from other pages at the site, in the way that Thelwall's specialised webometric spider (Thelwall 2000) does. However in comparison with e-journals, a positive correlation (Excel correlation coefficient = 0.45) can be made between the external WIF and the extent of total metadata use. There is a negative correlation (Excel correlation coefficient = -0.21) with the extent of DC metadata use, but this may be because of the small amount of DC metadata in the sample. The proportion of pages with metadata were similar to those for the e-journals.

| | |
|---|---|
| % pages with metadata | 16.86 |
| % pages with DC metadata | 4.35 |

## Discussion

Perhaps most significant was the relatively small amount of metadata use found in the study. Even in Library and Information Management e-journals, metadata was relatively rare; to the extent that at least one article on the topic of metadata did not include metadata in the HTML header (Caplan 1995).

Is use of metadata increasing? The increase between the Lawrence and Giles figures metadata (Lawrence & Giles 1999) and those found in this study are encouraging; however use of metadata by university sites and electronic journals would be expected to be higher than the norm. On the other hand, perhaps we don't want metadata to be too widely used: to some extent metadata acts as a filter, so that material that is worth retrieving will have metadata added, while more transitory material will not have metadata attached.

The study does demonstrate that the amount of metadata attached to a site influences at least some measures of the impact of a site. The correlation between the amount of Dublin Core metadata in elec-

tronic journal sites and the substantive external WIF is the most positive. For electronic journals, there is a slight negative correlation between the amount of metadata use and the standard external WIF; this may indicate the lack of validity of the standard external WIF as an impact measure for electronic journals, since this measure does not distinguish between links to the electronic journal as an entity, and links to substantive content. For NZ University sites, there is a positive correlation between the total metadata use and the impact factor of the site.

While these results are mixed, they are encouraging, given the effort expended on defining metadata standards. We may be approaching a critical mass of metadata, where metadata is sufficiently widely used in certain contexts to achieve usefulness, and will be adopted by search engines. According to Sullivan (Sullivan 2002), meta description tags are utilised by all major search engines except Google; meta keyword tags are utilised by Altavista and Inktomi, but not by FAST and Google.

This preliminary research does not positively confirm the value or otherwise of metadata. It indicates the need for further research to confirm the results of this exploratory study. In particular larger samples could be used to confirm the extent of metadata use by target sites. Larger numbers, particularly of university/research sites, and other classes of sites could be studied. The effect of quality and quantity of metadata used could also be studied.

## References

[URLs checked 13 June 2002].

Almind, T.C. & Ingwersen, P. 1997, 'Informetric analyses on the World Wide Web: methodological approaches to "Webometrics"', *Journal of Documentation,* vol. 53, no. 4, p. 404–426.

Björneborn, L. 2002, 'Defining webometrics [Message to webometrics@coombs.anu.edu.au 30 May 2002]'.

Brin, S. & Page, L. 1998, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Available: [http://www-db.stanford.edu/~backrub/google.html].

Caplan, P. 1995, 'You Call It Corn, We Call It Syntax-Independent Metadata for Document-Like Objects', *Public-Access Computer Systems Review,* vol. 6, no. 4. Available: http://info.lib.uh.edu/pr/v6/n4/capl6n4.html

Henshaw, R. & Valauskas, E. J. 2001, 'Metadata as a catalyst: experiments with metadata and search engines in the Internet journal, First Monday.', *Libri,* vol. 51, no. 2, pp. 86-101.

Ingwersen, P. 1998, 'Web Impact Factors', *Journal of Documentation,* vol. 54, no. 2, p. 236–243.

Lawrence, S. & Giles, C.L. 1999, 'Accessibility of information on the Web', *Nature,* vol. 400, pp. 107-9. Available: summary at http://www.wwwmetrics.com/.

Oppenheim, C., Morris, A., McKnight, C. & Lowley, S. 2000, 'The evaluation of WWW search engines', *Journal of Documentation,* vol. 56, no. 2, pp. 190-211.

Smith, A.G. 1999, 'A tale of two web spaces: comparing sites using web impact factors', *Journal of Documentation,* vol. 55, no. 5, pp. 577-92.

Smith, A.G. & Thelwall, M. 2002, 'Web Impact Factors for Australasian universities', *Scientometrics,* vol. 54, no. 3, p. 363–380.

Sullivan, D. 2002, *Search Engine Features For Webmasters*, Available: [http://www.searchenginewatch.com/webmasters/features.html].

Thelwall, M. 2000, 'Extracting Macroscopic information from web links', *Journal of the American Society for Information Science & Technology,* vol. 52, no. 13, pp. 1157-1168.

## Appendix 1. Electronic Journals

| Name | URL | Pages [P] | External links [X] | % metadata (keyword/ description) | % DC metadata | % total metadata | % links to articles [L] | WIF | Substantive WIF |
|---|---|---|---|---|---|---|---|---|---|
| LibRes: Library And Information Science Research Electronic Journal | libres.curtin.edu.au | 38 | 67 | 0 | 0 | 0 | 0 | 1.76 | 0.00 |
| Electronic Journal of Radical Organisation Theory | mngt.waikato.ac.nz/ejrot | 46 | 122 | 0 | 0 | 0 | 0 | 2.65 | 0.00 |
| Canadian Journal of Educational Administration and Policy (CJEAP) | umanitoba.ca/publications/cjeap/ | 22 | 128 | 0 | 0 | 0 | 0 | 5.82 | 0.00 |
| Electronic Journal of Probability | math.washington.edu/~ejpecp | 130 | 982 | 0 | 0 | 0 | 0 | 7.55 | 0.00 |
| Electronic Transactions on Numerical Analysis (ETNA) | etna.mcs.kent.edu | 87 | 977 | 0 | 0 | 0 | 0 | 11.23 | 0.00 |
| Earth Interactions: An Electronic Journal Serving the Earth System Science Community | earthinteractions.org | 12 | 279 | 0 | 0 | 0 | 0 | 23.25 | 0.00 |
| EM Electronic Markets | electronicmarkets.org | 605 | 934 | 100 | 0 | 100 | 0 | 1.54 | 0.00 |
| Cybermetrics | cindoc.csic.es/cybermetrics | 81 | 288 | 100 | 0 | 100 | 0 | 3.56 | 0.00 |
| Architronic | architronic.saed.kent.edu | 292 | 110 | 0 | 0 | 0 | 20 | 0.38 | 0.08 |
| Journal of Information Law and Technology | elj.warwick.ac.uk/jilt | 2,325 | 1,298 | 0 | 0 | 0 | 30 | 0.56 | 0.17 |
| Electronic Journal of Biotechnology | ejb.org | 245 | 417 | 80 | 0 | 80 | 10 | 1.70 | 0.17 |
| E Law - Murdoch University Electronic Journal of Law | murdoch.edu.au/elaw | 1545 | 972 | 0 | 0 | 0 | 30 | 0.63 | 0.19 |
| Information Research | informationr.net/ir | 119 | 230 | 20 | 30 | 50 | 10 | 1.93 | 0.19 |
| Crossings Electronic Journal of Art and Technology | crossings.tcd.ie | 22 | 23 | 0 | 0 | 0 | 30 | 1.05 | 0.31 |
| Electronic musicological review | cce.ufpr.br/~rem | 189 | 307 | 0 | 0 | 0 | 20 | 1.62 | 0.32 |
| Folklore: An Electronic Journal of Folklore | haldjas.folklore.ee/folklore | 462 | 518 | 0 | 0 | 0 | 40 | 1.12 | 0.45 |
| International Electronic Journal for Leadership in Learning | ucalgary.ca/~iejll | 71 | 435 | 0 | 0 | 0 | 10 | 6.13 | 0.61 |
| Reading Online | readingonline.org | 805 | 1697 | 40 | 0 | 40 | 30 | 2.11 | 0.63 |
| E-Journal | hanover.edu/philos/ejournal | 140 | 150 | 0 | 0 | 0 | 60 | 1.07 | 0.64 |
| Interactive multimedia electronic journal of computer-enhanced learning | imej.wfu.edu | 247 | 356 | 20 | 0 | 20 | 50 | 1.44 | 0.72 |
| Australasian Journal of Disaster and Trauma Studies | massey.ac.nz/~trauma | 76 | 279 | 10 | 0 | 10 | 20 | 3.67 | 0.73 |
| Screening the Past: An International Electronic Journal of Visual Media and History | latrobe.edu.au/www/screeningthepast | 349 | 855 | 0 | 0 | 0 | 40 | 2.45 | 0.98 |

## Appendix 1. Electronic Journals (continued)

| Name | URL | Pages [P] | External links [X] | % metadata (keyword/ description) | % DC metadata | % total metadata | % links to articles [L] | WIF | Substantive WIF |
|---|---|---|---|---|---|---|---|---|---|
| Journal of World-Systems Research | csf.colorado.edu/jwsr | 51 | 259 | 10 | 0 | 10 | 20 | 5.08 | 1.02 |
| Ariadne | ariadne.ac.uk | 1073 | 2556 | 0 | 10 | 10 | 50 | 2.38 | 1.19 |
| Journal of Digital Information | jodi.ecs.soton.ac.uk | 270 | 880 | 40 | 0 | 40 | 40 | 3.26 | 1.30 |
| The Journal of Library Services for Distance Education | westga.edu/~library/jlsde | 40 | 523 | 0 | 0 | 0 | 10 | 13.08 | 1.31 |
| Interpersonal Computing and Technology; An Electronic Journal for the 21st Century | jan.ucc.nau.edu/~ipct-j | 18 | 404 | 0 | 0 | 0 | 10 | 22.44 | 2.24 |
| Journal of Electronic Publishing | press.umich.edu/jep | 244 | 2,007 | 20 | 0 | 20 | 30 | 8.23 | 2.47 |
| Dlib Magazine | dlib.org/dlib | 882 | 3581 | 20 | 10 | 30 | 80 | 4.06 | 3.25 |
| Essays in History | etext.lib.virginia.edu/journals/EH | 69 | 411 | 0 | 0 | 0 | 60 | 5.96 | 3.57 |
| Journal of Computer-Mediated Communication | ascusc.org/jcmc | 322 | 2524 | 70 | 0 | 70 | 60 | 7.84 | 4.70 |
| First Monday | firstmonday.org | 157 | 2030 | 0 | 60 | 60 | 50 | 12.93 | 6.46 |
| Public-Access Computer Systems Review (PACS Review) | info.lib.uh.edu/pr | 49 | 920 | 0 | 0 | 0 | 90 | 18.78 | 16.90 |

## Appendix 2. NZ University websites

| Name | URL | pages | external links | % metadata (keyword/ description) | % DC metadata | % total metadata | WIF |
|---|---|---|---|---|---|---|---|
| Massey | massey.ac.nz | 213,151 | 10,824 | 0 | 0 | 0 | 0.05 |
| Auckland Univ of Technology | aut.ac.nz | 10,355 | 2,112 | 10 | 0 | 10 | 0.20 |
| Otago | otago.ac.nz | 32,662 | 11,039 | 80 | 0 | 80 | 0.34 |
| Waikato | waikato.ac.nz | 32,639 | 12,251 | 10 | 30 | 40 | 0.38 |
| Canterbury | canterbury.ac.nz | 32,408 | 13,588 | 20 | 10 | 30 | 0.42 |
| Auckland | auckland.ac.nz | 35,431 | 21691 | 0 | 10 | 10 | 0.61 |
| Lincoln | lincoln.ac.nz | 3,011 | 3,936 | 70 | 0 | 70 | 1.31 |
| Victoria | vuw.ac.nz | 21,781 | 31,303 | 40 | 0 | 40 | 1.44 |

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 139-146*


# Using Dublin Core to Build a Common Data Architecture

Sandra Fricker Hostetter
Rohm and Haas Company, Knowledge Center
shostetter@rohmhaas.com

## Abstract

*The corporate world is drowning in disparate data. Data elements, field names, column names, row names, labels, metatags, etc. seem to reproduce at whim. Librarians have been battling data disparity for over a century with tools like controlled vocabularies and classification schemes. Data Administrators have been waging their own war using data dictionaries and naming conventions. Both camps have had limited success. A common data architecture bridges the gap between the worlds of tabular (structured) and non-tabular (unstructured) data to provide a total solution and clear understanding of all data. Using the Dublin Core Metadata Element Set Version 1.1 and its Information Resource concept as building blocks, the Rohm and Haas Company Knowledge Center has created a common data architecture for use in the implementation of an electronic document management system (EDMS). This platform independent framework, when fully implemented, will provide the ability to create specific subsets of enterprise data on demand, enable interoperability with other internal or external systems, and reduce cycle time when migrating to the next generation tool.*
**Keywords:** *common data architecture, CDA, document management, platform independent framework, data resource management, metadata, Dublin Core, controlled vocabularies*

## 1. A new hybrid

Organizing information has become a core competency for corporations. Moving from a paper-based world to an electronic-based one is a difficult and lengthy transformation. Paper forced us to behave in certain ways because of physical limitations associated with its tangibility. However, paper also had inherent strengths in its universality and this is something we have taken for granted.

Blending the features of paper and electronic formats is an enormous challenge. We must create something new. The plant world provides us with a helpful analogy. A hybrid plant is the combination of two separate entities into something completely new and unique, yet shares the attributes of both parent plants. This does not happen by accident. Two different species of plants will not merge to create a new one without purposeful human intervention, management, and care. And therein lie both the problem and the opportunity.

In the past, tabular and non-tabular data have been managed and accessed in very different ways. However, the ever-demanding user population wants to see all the available data integrated together and presented in a manner individually tailored to their specific needs. It has become impossible to separately manage non-tabular data and tabular data. This demands we address seemingly mutually exclusive issues in a way that satisfies all parties. The creation of a common data architecture is the most effective way to bridge the gap between all types of data.

## 2. Metadata management in a document managed world

The importance of controlling the metadata used to describe items deposited in a document management system is critical to facilitate effective search and retrieval activities in partnership with the dueling aspects of a full-text environment – instant gratification and lack of discrimination. At the Rohm and Haas Company, Dublin Core was a good starting point and became the basis for the document class and document properties structure "dictated" by the EDMS. From the beginning, our goal was to create a platform independent framework that would meet the following needs: (1) enable the creation of specific subsets of enterprise data on demand (2) provide future interoperability with other internal and external systems (3) reduce cycle time when migrating from "today's tool," to the next generation of document management software without excessive rework.

The Dublin Core data elements as implemented in the EDMS at the Rohm and Haas Company function as the common metadata. All document classes have these properties, though it is not mandatory the properties be populated. Eventually, three of these Dublin Core based properties (DC.Title, DC.Date.issued, DC.Publisher) will be required, and DC.Publisher will have a Rohm and Haas specific controlled scheme to reflect the company's business unit structure.

## 3. The common data architecture approach

A common data architecture (CDA) "is a formal, comprehensive, data architecture that provides a common context within which ALL DATA are understood and integrated". A CDA has the following basic components – data subjects, data characteristics, and data characteristic variations. A *data subject* is "a person, place, thing, concept, or event that is of interest to the organization and about which data are captured and maintained". A *data characteristic* is "an individual characteristic that describes a data subject". A *data characteristic variation* "represents a difference in the format, content, or meaning of a specific data characteristic" (Brackett, 1994, p. 31, p. 39).

At first glance, a standard like the Dublin Core Metadata Element Set Version 1.1 looks like it might be a common data architecture. However under closer scrutiny, its deficiencies become more obvious. Dublin Core violates a core principle of data management by mixing different facts within a single field. DC.Creator can represent a person or an organization. The ideal data management equation is 1 Fact = 1 Field. In Dublin Core's well-intended effort to be simple yet fully extensible, it is also very non-specific. This leads us down the tempting path to the never-ending crosswalk. Cross walking happens only at the physical level, requires an excessive amount of work, and yields minimal understanding. Instead, if we move beyond the traditional physical level analysis and cross-reference to a common data architecture created at the logical level, we gain a true common context for understanding all data.

## 4. How to build a common data architecture

Building a common data architecture involves five major steps. It is a reiterative process that may take several months to become an accurate reflection of the organizational situation and will require occasional readjustments over time. Since a common data architecture represents is a living breathing organization that grows and changes, it too must be refreshed as needed.

### 4.1 Defining the "pivotal" data subject

The first step is to identify, formally name, and define the pivotal data subject. The pivotal data subject is the most central business concept. All related concepts will be organized around this data subject. The pivotal data subject for the EDMS was the software defined object "Document Class". We adopted the Dublin Core terminology for "Information Resource" and broadened the definition as follows:

*Information Resource*

An Information Resource is a set of data in context, recorded in any medium of expression (text, audio, video, graphic, digital) that is meaningful, relevant, and understandable to one or more people at a point in time or for a period of time. Traditionally, an Information Resource is recorded on some medium, such as a document, a web page, a diagram, and so on. In the broad sense, however, an Information Resource could be a person or a team of people.

An Information Resource in this data architecture represents a version of an Information Resource when there is more than one version produced. The Information Resource. System Identifier changes for each version. The Information Resource Document. Number that is assigned as an Information Property Item through Information Resource Property remains the same across versions and identifies the Information Resource, and the Information Resource. Version Identifier uniquely identifies the version of that Information Resource.

Note that the system identifier as defined in this data architecture is the system identifier of the home system where data about information resources are stored. Any other foreign identifiers from other systems where data about information resources are stored are assigned as an Information Property Item through Information Resource Property.

Note that there are non-EDMS versions of an Information Resource, such as web page versions, that may not have a date, version identifier, URL change, and so on. There is no way to know or distinguish versions of this type.

### 4.2 Defining the data characteristics

The second step is to identify, formally name, and define the data characteristics of the pivotal data subject. Examples include:

*Information Resource. Title*

The official title of the Information Resource, such as "The Importance of Adding Property Data to a Panagon Document." This is the name by which the Information Resource is formally known.

*Information Resource. System Identifier*

The system assigned identifier in the home system that uniquely identifies an Information Resource. This is not the same as the system identifier that identifies an Information Resource in an EDMS system or any other foreign system documenting Information Resources. The Information Resource, System Identifier changes for each version of an Information Resource. The Information Resource. Version Identifier identifies the version of the Information Resource.

*Information Resource. Version Identifier*

The version number of the Information Resource. The versions are typically, but not necessarily, assigned sequentially from 1. In some foreign systems or standards, the version identifier may be appended to the system identifier. In this data architecture, the version identifier is maintained separate from the system identifier.

*Information Resource Subtype. Code*

Information Resource Subtype indicates a more detailed classification of documents within Information Resource Type. Not every Information Resource Type will have Information Resource Subtypes.

*Information Resource Type. Code*

The code that unqiuely identifies an Information Resource Type, such as LNBK for the Information Resource Type Laboratory Notebook.

### 4.3 Defining the qualifying data subjects

The third step is to identify, formally name, and define any qualifying data subjects and their data characteristics. We used the Dublin Core Metadata Element Set Version 1.1 as the basic building blocks. Examples include:

*Information Contributor*

An Information Contributor is any person or organization that contributes in any way to an Information Resource. A person may be an author, a researcher that provides material, or a reviewer, and an organization may be a service or professional organization. Information Resource Contributor connects an Information Contributor to an Information Resource. Information Resource Contributor Role identifies the specific role played by an Information Contributor.

*Information Property Group*

An Information Property Group is a set of related Information Property Items. The structure of Information Property Groups and Information Property Items allows a variety of reference tables or enumerated lists to be defined for assignment to an Information Resource through Information Resource Property. Information Property Group represents a controlled set of reference tables

*Information Property Item*

Information Property Item is one reference item from a set of reference items commonly held by an Information Resource. Each Information Property Item belongs to an Information Property Group. Information Resource Property assigns the Information Property Items to Information Resources.

*Information Property Item Alias*

An Information Property Item can have different names in different systems or standards. There is no uniform name that transcends all systems and standards. Information Property Item Alias documents all of the alias names for a foreign Information Property Items in various systems and standards, and their originating system or standard. The preferred name is shown in Information Property Item. Name.

*Information Resource Contributor*

An Information Resource can have many different Information Contributors, and an Information Contributor can contribute to many different Information Resources. Information Resource Contributor designates a specific Information Contributor for a specific Information Resource. Information Resource Contributor Role identifies the specific role performed by an Information Resource Contributor.

*Information Resource Contributor Role*

An Information Contributor can perform different roles with respect to an Information Resource. Information Resource Contributor Role is a reference table identifying the roles that an Information Contributor can perform for an Information Resource.

*Information Resource Property*

An Information Resource can be characterized by many different Information Property Items, and an Information Property Item can characterize many different Information Resources. Information Resource Property assigns a specific Information Property Item to a specific Information Resource. If that Information Property Item requires additional data, such as a date or description, those data are provided in the data characteristics described below.

*Information Resource Property Validity*

An Information Resource Type has a set of Information Properties Items that are valid and can be assigned to an Information Resource belonging to that Information Resource Type. Information Resource Property Validity indicates the valid assignments of Information Property Items. Note that this data subject is set up to show only the valid assignments of an Information Property Item for an Information Resource Type. If an Information Property Item appears, then that Information Property Item is valid for the Information Resource Type. If an Information Property Item does not appear, then that Information Property Item is not valid for the Information Resource Type.

*Information Resource Publisher*

An Information Resource can be published by more than one Publisher, and a Publisher can publish more than one Information Resource. Information Resource Publisher identifies the publication of an Information Resource by a specific Publisher.

*Information Resource Relationship*

An Information Resource can have a relationship with other Information Resources, such as reference to another Information Resource, material included from another Information Resource, and so on. Information Resource Relationship identifies a specific relationship between two Information Resources. Information Resource Relationship Type identifies the specific type of relationship between Information Resources.

*Information Resource Relationship Type*

Information Resource Relationship Type is a reference table that identifies the specific type of relationship between two Information Resources identified in Information Resource Relationship.

*Information Resource Subtype*

Information Resource Subtype indicates a more detailed classification of documents within Information Resource Type. Not every Information Resource Type will have Information Resource Subtypes.

*Information Resource Type*

Information Resource Type is a broad grouping of Information Resources that designates the nature or genre of the content of the Information Resource. It describes general categories, functions, or aggregation levels of the content of Information Resources.

*Information Security Group*

An Information Resource can have different levels of security classification governing which individuals or organizations can access that Information Resource. Information Security Group is a reference table designating the broad levels of security for an Information Resource. Information Security Subgroup identifies a more detailed grouping of security.

*Information Security Subgroup*

Information Security Groups can have a more detailed level of classification. Information Security Subgroup provides the detailed levels of security classification within Information Security Group.

*Publisher*

A Publisher is any organization, internal or external to Rohm and Haas, that formally publishes an Information Resource. Note that this current definition is limited to Information Resources. As the common data architecture is enhanced, this definition may be altered to include the publishers of other material not considered an Information Resource.

**4.4 Creating a visual representation of the relationships**

The fourth step is to create a visual representation of how all the data subjects relate to each other. In Figure 1 the relationships are depicted in a manner based on data modeling techniques outlined below:

Arrows moving away from a data subject represent a one-to-many relationship between the data subjects. For example, a single Information Resource may have many Information Resource Contributors. An Information Contributor (DC.Creator or DC.Contributor) is any person or organization that contributes in any way to an Information Resource. A person may be an author, a researcher that provides material, or a reviewer, and an organization may be a service or a professional organization

Arrows moving towards a data subject represent a many-to-one relationship. For example, an Information Contributor may be an Information Resource Contributor to many different Information Resources. Information Resource Contributor designates a specific Information Contributor for a specific Information Resource. Information Resource Contributor Role identifies the specific role performed by an Information Resource Contributor.

Two arrows represent a relationship between two Information Resources. For example, an Information Resource can have a relationship with other Information Resources, such as reference to another Information Resource, material included from another Information Resource, etc. Information Resource Relationship identifies a specific relationship between two Information Resources. Information Resource Relationship Type identifies the specific type of relationship between Information Resources.

Multiple arrows going in the same direction in sequence represent a hierarchy relationship. For example, Information Resource Subtype indicates a more detailed classification of documents within Information Resource Type. However, not every Information Resource Type will have Information Resource Subtypes.

Arrows going towards each other and intersecting at the same data subject represent an assignment relationship. For example, Information Resource Contributor connects an Information Contributor to an Information Resource. Information Resource Contributor Role identifies the specific role played by an Information Contributor (the various Information Resource Contributor Roles that an Information Contributor can perform for an Information Resource are stored in a reference table. This will be discussed in more detail under heading 5. Properties Make the World Go 'Round). We assign an Information Contributor to an Information Resource and then we give the Information Contributor a specific role. The same kind of assignment relationship exists for Publisher. An Information Resource could be published by two different publishers. The print copy could be pub-

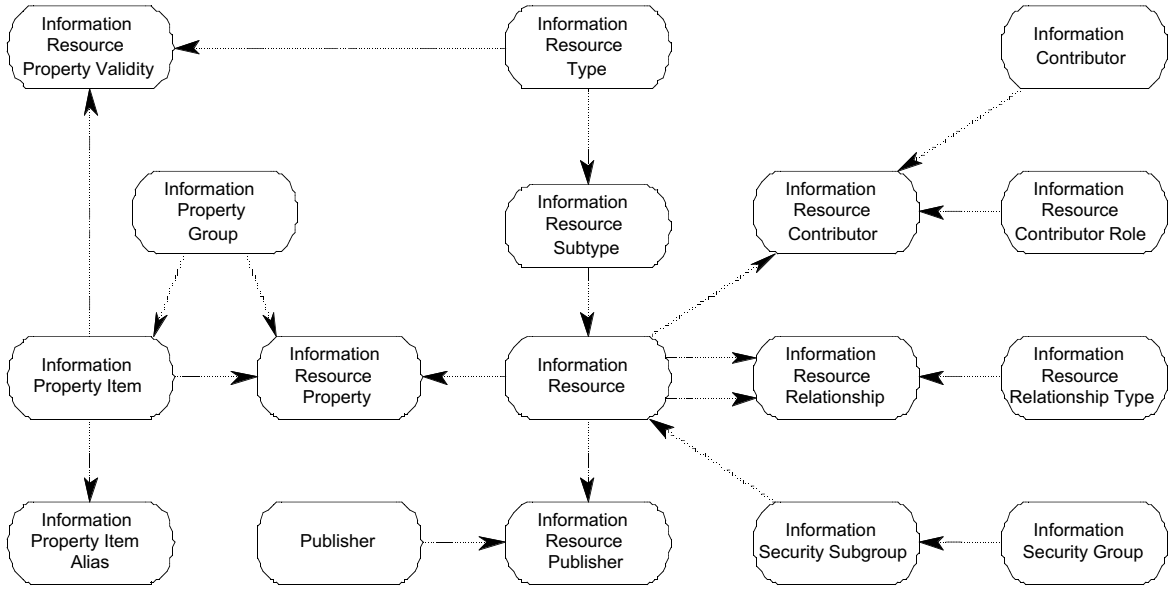



**Figure 1.**

lished by a different entity than the electronic version and the electronic and print versions could be the same content or might be different content.

#### 4.5 Testing and adjustment

The final step is to test the resulting common data architecture and adjust as needed. This can be done by trying it out on another system or conducting business use cases.

## 5. Properties make the data world go 'round

Properties (fields, attributes, characteristics, features, metatags) help us understand more about the content and context of the information resource to which they belong. Common properties are universal. Everyone in the organization cares about these properties. It is important to limit the names and display labels of these common properties so we can effectively share them and mean the same thing. Special or custom properties apply only to a small subset of information resources, but their names and labels should be limited also. Limiting the values for most properties helps keep the context meaningful and clear.

Because an Information Resource may have many different Information Resource Property Items, we need to resolve the many-to-many relationship and figure out a way to assign them to the specific Information Resource. We define the Information Resource Property Items first, and then assign them. By structuring things in this manner, Information Property Groups and Information Property Items within those groups can become ineffective at any time without altering the structure of the data resource (Figure 2).

Information Resource Property Items for a specific Information Resource are kept in reference tables called Information Property Groups. An Information Resource Property is a qualifying Data Subject which assigns Information Resource Property Items, via the Information Resource Property Groups structure to a specific Information Resource.

Information Property Group is a reference table of reference tables. Information Property Item is a specific value in a reference table. All Information Property Items must belong to an Information Property Group. This portion of the CDA represents "a controlled vocabulary of controlled vocabularies". These reference tables are documented as data subjects, but their definitions clearly identify them as reference tables and not true data subjects.

An example of an Information Resource Property

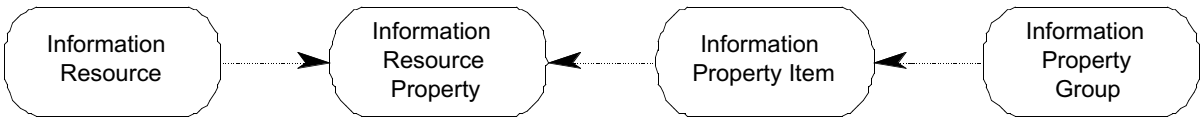



**Figure 2.**

Group is Information Resource Description. An Information Resource can have many associated descriptions, such as content, spatial, physical format, temporal, and rights. Information Resource Description identifies each of the descriptions that can be assigned to an Information Resource. Examples of Information Property Items for the Information Resource Property Group called "Information Resource Description" are Content Description, Spatial Description, Physical Format Description, Temporal Description, and Rights Description. Other examples of Information Property Groups include: Information Resource Date, Information Resource Identifier, Information Resource Library, Information Resource Subject, Language, and Non-Enumerated Feature.

## 6. Documenting the common data architecture

We are formally documenting our common data architecture in the Data Resource Guide. The Data Resource Guide is a proprietary Microsoft Access software application which contains tables on the Common Data Architecture side for data subject, data characteristic, data characteristic variation, data code set, data code. On the Data Product side (e.g. EDMS, Dublin Core Metadata Element Set Version 1.1, etc.) the database has tables for data product type, data product, data product group, data product unit, and data product code. It also has tables for data product cross-referencing. The inclusion of a reporting feature enables the data resource administrator to see how multiple data products relate to each other and what data elements they share.

## 7. Cross referencing Dublin Core to the CDA

When we cross-reference the Dublin Core Metadata Element Set Version 1.1 to the common data architecture it yields the following results.

| Dublin Core Element Label | Common Data Architecture Equivalent (Data Subject. Data Characteristic, Data Characteristic Variation) |
|---|---|
| Title | **Information Resource. Title, Variable** |
| Creator | Creator can be either a person or an organization. The cross-references are identified for each Creator variant.<br>**Information Contributor. Person Name, Complete Inverted**<br>Comment: Information Resource Contributor Role. Name, Formal = 'Creator'<br>**Information Contributor. Organization Name, Variable**<br>Comment: Information Resource Contributor Role. Name, Formal='Creator' |
| Subject | An exact cross-reference is indeterminate based on the definition of Subject and the lack of a specific controlled vocabulary or formal classification scheme. Any implementation could use one or more controlled vocabularies or formal classification schemes. The best cross-reference approach is to identify each specific controlled vocabulary or formal classification scheme used under the Dublin Core standard, document it as a reference table in the common data architecture, and then prepare a cross-reference to that reference table.<br>**Business Unit Classification Scheme. Name, Formal**<br>Comment: Information Property Group. Name, Formal = Information Resource Subject<br>Comment: Information Property Group. Name, Formal is indeterminate and needs to be determined for each data occurrence. |
| Description | Description is defined as a reference table in the common data architecture as Information Resource Description. The specific types of descriptions, such as table of contents, abstract, etc. are reference items in that reference table.<br>**Information Resource Property. Description, Dublin Core**<br>Comment: Information Property Group = Information Resource Description<br>Comment: Information Property Item. Name, Formal is variable and needs to be determined for each data occurrence. |
| Publisher | **Publisher. Name, Variable**<br>Comment: The publisher name should be used as the cross-reference. |
| Contributor | Contributor can be either a person or an organization. The cross-references are identified for each Contributor variant.<br>**Information Contributor. Person Name, Complete Inverted**<br>Comment: Information Resource Contributor Role. Name, Formal is variable and needs to be determined for each data occurrence.<br>**Information Contributor. Organization Name, Variable**<br>Comment: Information Resource Contributor Role. Name, Formal is variable and needs to be determined for each data occurrence. |
| Date | Date is defined as a reference table in the common data architecture as Information Resource Date. The specific types of dates, such as Available Date, Creation Date, Issued Date, Modified Date, Valid Date, etc. are reference items in that reference table.<br>**Information Resource Property. Date, ISO 8601** |

Comment: Information Property Group. Name, Formal=Information Resource Date
Comment: Information Property Item. Names, Formal is variable and needs to be determined for each data occurrence.

Type **Information Resource Type, Name, Dublin Core**
Comment: If a controlled vocabulary other than the list of Dublin Core types is used, it needs to be documented as a data product unit variant and cross-referenced to an appropriate reference table in the common data architecture.

Format Format is a specific type of description which is defined as a reference table in the common data architecture as Information Resource Description. The specific types of descriptions, such as text, audio, etc. are reference items in that reference table.
**Information Resource Property. Description, Dublin Core**
Comment: Information Property Group. Name, Formal = Information Resource Description
Comment: Information Resource Description. Name, Formal is variable and needs to be determined for each data occurrence.

Identifier Identifier is defined as a reference table in the common data architecture as Information Resource Identifier. The specific types of identifiers, such as URI, ISBN, etc., are reference items in that reference table.
**Information Resource Property. Value, Variable**
Comment: Information Resource Property = Information Resource Identifier
Comment: Resource Description. Name, Formal is variable and needs to be determined for each data occurrence

Source Source represents a relationship between two Information Resources as defined in Information Resource Relationship. The identifier of the Source in Dublin Core must be determined, the system identifier located, and that system identifier used in Information Resource Relationship. System Identifier. The specific types of relationships, such as source, and so on, are defined in Information Resource Reference Type.
**Information Resource Property. Value, Identifier Variable**
Comment: Information Resource Relationship Type. Name, Formal = Source

Language Language is a multiple-fact data item for the language and the country associated with the language. The cross-references are identified for each language variant.
**Language. Code, ISO 639**
**Country. Code, ISO 3166**

Relation Relation represents a relationship between two Information Resources as defined in Information Resource Relationship. The identifier of the Source in Dublin Core must be determined, the system identifier located, and that system identifier used in Information Resource Relationship. System Identifier. The specific types of relationships, such as source, etc. are defined in Information Resource Reference Type.
**Information Resource Property. Value, Identifier Variable**
Comment: The Information Resource Relationship Type. Name, Formal is indeterminate and needs to be identified for each data occurrence.

Coverage Coverage is a specific type of description which is defined as a reference table in the common data architecture as Information Resource Description. The specific types of coverage, such as spatial, temporal, etc. are reference items in that reference table. Information Resource Description. Name is variable and needs to be determined for each data occurrence.
**Information Resource Property. Description, Dublin Core**
Comment: Information Resource Property. Name, Formal = Description
Comment: Information Property Item. Name, Formal = Spatial Description
Comment: Information Property Item. Name, Formal = Temporal Description
Comment: Information Property Item. Name, Formal = Jurisdiction Description

Rights Rights is a specific type of description which is defined as a reference table in the common data architecture as Information Resource Description. The specific types of rights, such as copyright, royalty, and so on, are reference items in that reference table.
**Information Resource Property. Description, Dublin Core**
Comment: Information Property Group. Name, Formal = Information Resource Description
Comment: Information Property Item. Name, Formal is variable and needs to be deter mined for each data occurrence.

## 8. Next steps

This common data architecture is currently a work-in-progress. Full documentation of the common data architecture in a Data Resource Guide must be completed as well as the final cross-referencing of the EDMS metadata and Dublin Core Metadata Element Set 1.1. The creation of a thesaurus component is essential to making the CDA content available to the wider community of general

system users and the individuals who develop and design new database applications. We envision a "Data Element Supermarket" where developers can shop for the field name desired, find its variations (code, name, acronym), and learn its single source and history of use in other systems. We have created a good foundation, but there is still much work to be done before the true value can be realized.

## Acknowledgements

This work would not have been possible without the professional advice and consultation expertise of Michael H. Brackett. The author is grateful for his personal encouragement and support.

## References

1. Brackett, M., 1994. *Data Sharing Using a Common Data Architecture*. New York: John Wiley & Sons, Inc.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 147-156*


# Using Web Services to Interoperate Data at the FAO

Andrea Zisman
Department of Computing
City University
Northampton Square
London EC1V 0HB,UK
a.zisman@soi.city.ac.uk

John Chelsom, Niki Dinsey
CSW Informatics Ltd
4240 Nash Court
Oxford Business Park South
Oxford OX4 2RU, UK
john.chelsom@csw.co.uk
niki.dinsey@csw.co.uk

Stephen Katz, Fernando Servan
Food and Agriculture Organization
United Nations - WAICENT
Viale delle Terme di Caracalla
Rome 00100, Italy
stephen.katz@fao.org fernando.servan@fao.org

## Abstract

*In this paper we present our experience of using Web services to support interoperability of data sources at the Food and Agriculture Organization of the United Nations. We describe the information bus architecture based on Web services to assist with multilingual access of data stored in various data sources and dynamic report generation. The architecture preserves the autonomy of the participating data sources and allows evolution of the system by adding and removing data sources. In addition, due to the characteristics of Web services of hiding implementation details of the services, and therefore, being able to be used independently of the hardware or software platform in which it is implemented, the proposed architecture supports the problem of existing different technologies widespread in the FAO, and alleviates the difficulty of imposing a single technology throughout the organization. We discuss the benefits and drawbacks of our approach and the experience gained during the development of our architecture.*
**Keywords:** *XML, Web services, J2EE, .NET, Ontologies, RDF, Topic Maps, WAICENT.*

## 1. Introduction

The development of distributed computing and networking has provided the technological basis for remote access to data and applications. The development of different systems has increased the utility of these systems, but has not solved the problem of having a large number of applications interoperating with each other. The applications have not been built to be integrated, and therefore, they normally define different data formats, have their own communication protocols, and are developed on different platforms. Interoperability of distributed systems is still a challenge.

Nowadays it is important to allow interoperability of different types of information sources in a large company or community. Users and applications have a growing need to access and manipulate data from a wide variety of information sources. However, the data sources are generally created and administered independently, differing physically and logically. Other difficulties associated with such an environment include: heterogeneity and autonomy of database systems, conflict identification and resolution,

semantic representation of data, location and identification of relevant information, access and unification of remote data, query processing, and easy evolution of the system.

An example of the above problem is found in the Food and Agriculture Organization of the United Nations (FAO). FAO is a specialized agency of the United Nations, which leads international efforts to defeat hunger. It helps developing countries modernize and expand agriculture, forestry and fisheries and ensure good nutrition for all. One of its most important functions is to collect, analyze and disseminate information to assist governments to fight hunger and achieve food security. Towards this effort FAO has established the World Agricultural Information Centre (WAICENT) for agricultural information management and dissemination.

Within the WAICENT framework, a large amount of data, represented in various distinct formats, in many different languages, and handled by several metadata structures, are generated every day and stored in different types of data sources. However, there are no standards for representing languages, metadata, and specific country information. People need to access and manipulate data distributed in the various sources from both inside and outside the organization. It is important to share data between systems quickly and easily, without requiring the systems to be tightly coupled. In simple terms, the existing systems need to "talk" to each other. Another main problem is related to the fact that within the organization the use of two different technologies (Microsoft ASP [5] and Java JSP/servlets [20]) is widespread and it is, therefore, very difficult to impose a single technology throughout the FAO.

In this paper we present an approach based on Web services [17] and eXtensible Markup Language (XML) [6] technology to allow interoperability of the different data sources in the FAO. It is a lightweight approach and is based on the use of an *information bus* to allow exchanged of data between various information sources implemented by using different technologies. The *information bus* supports multilingual access of data stored in various data sources, handles metadata in a generic way, and enables metadata to be used as exchange models throughout FAO. The approach also supports dynamic report generation. A prototype tool has been implemented to demonstrate and evaluate the approach.

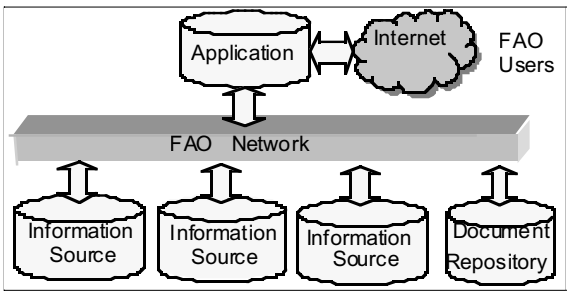



**Figure 1. Existing information structure at FAO**

The remaining of the paper is organized as follows. Section 2 describes the problem in the FAO that is being tackled by our approach. Section 3 presents some related work. Section 4 outlines the *information bus* and the dynamic report generation. Section 5 illustrates our work through examples. Section 6 discusses the implementation of our prototype and evaluation issues. Finally, section 7 summarizes our experience and suggests directions for future work.

## 2. The problem

The Food and Agriculture Organization of the United Nations has approximately 200 systems supplying information for access on the World Wide Web, deployed using two different technologies: Microsoft ASP [5] and Java JSP/servlets [20]. These data sources need to share and exchange data between each other in an easy way. However, the use of the two technologies is already widespread in the organization and it is almost impossible to impose a single technology throughout the FAO. In addition, it is necessary to avoid rewriting of existing applications.

The existing information infrastructure is shown in Figure 1. It consists of information sources (database systems) containing different types of data including, but not limited to, different types of documents written in five official languages - English, French, Spanish, Chinese and Arabic (and some in Russian); electronic bibliography references; statistical data; maps and graphics; news and events from different countries; and web information.

Different people generate documents in different formats, which are inserted in the databases using web interfaces. The data is accessed from the databases in HTML format, through applications available on the Internet. Examples of these applications are WAICENT Information Finder (an online search tools), FAOBIB (an online catalogue of bibliography), FAO Virtual Library (a digital archive), and FAOSTAT (an online database about statistics of various areas). The FAO users are farmers, scientists, traders, government planners, and non-governmental people, both inside and outside the organization, that need to access and publish information.

Although the existing setting addresses some of the requirements of integrating disparate distributed systems, there are limitations involving budgetary or technical challenges, inflexibility, lack of standardization, and difficulty of scalability and extensibility. It is important to have a technology that is inexpensive, easy to implement, easy to maintain and based on open standards, to allow leverage of knowledge and

existing resources without having to rewrite existing applications.

The technology needs to support interoperability of existing data sources and management of multilingual variants without changing the database structures. Currently, it is necessary to customize and add database structures for each different language. There is no standard way to manage language variants of documents or other data structures. This generates inconsistencies between applications in the way that they manage the different languages. In addition, the database models are not easily extensible when new data or language variants are added.

Other problems were related to the support of metadata representation and metadata exchange in a standard way, as well as use of standard ontology formats. In the FAO a document repository has been developed with the objective of storing and disseminating all publications electronically. It stores meeting notes, documents, metadata, and index data.

Different ASP interfaces have been created to allow searching the document repository by type, language, and subject. However, there is no standard way to manage language variants of documents or other data structures like specific country information and metadata. The multilingual Agricultural Thesaurus (AGROVOC) [2] from FAO has been applied to the web as a strategy to ensure some conformity in resource description/discovery. However, it falls short of being a complete tool for this purpose in view of a need for more specific subject terminology and richer ontological relations that are offered by traditional thesaurus.

## 3. Related work

The challenge of interoperating distributed systems, in particular database systems, has existed for a long time and has been extensively researched. Many approaches have been proposed to allow integration and interoperability of distributed systems developed in an independent way. These approaches have been proposed as outcomes of research work in both academia and industry.

We can divide the existing approaches into two main groups [34]. In the first group of approaches a global schema is used as another layer on the top of existing schemas which gives the users and applications the illusion of a single, centralized database system. Examples of these approaches include systems like DATAPLEX [9], DDTS [13], MULTIBASE [29], and PEGASUS [3]. However, the construction of a global schema is not a simple task, does not guarantee the autonomy of the participating database systems, and does not allow easy evolution of the system in terms of adding and removing of participating databases.

In order to overcome the problem of constructing a global integrated schema the second group of approaches has been proposed, in which 'partial' or 'no integration' is performed. Examples of these approaches include the federated architecture [19], five-level schema architecture [28], multidatabase architecture [22][23], the Jupiter system [18], and [33]. Within the approaches that do not use a global schema some of them proposed the use of mediators and wrappers. In these approaches data sources are encapsulated to make it usable in a more convenient manner by hiding or exposing the internal interface of the data sources, reformat data, and translate queries. Examples of systems that use wrappers and mediators are DIOM [24], DISCO [31], Garlic [27], and TSIMMIS [16].

In any of the above approaches and existing technologies the problems related to how to format data to be exchanged and how to transmit the data are still open problems. Regarding data format, there are almost no tools that can automate the process of translating data in different formats. Many systems use ASCII-based text files to represent their data. However, there is no standard way of formatting or describing the values in the files. The different systems exchanging data in ASCII format must have custom-built loading software to handle different file formats. Other systems exchange data via a specified file format, which does not scale well (e.g. Microsoft Excel).

On the other hand, data transmission has also been difficult to implement. The use of the File Transfer Protocol (FTP) facilitates file transfer, but this is not a tight, object-oriented approach to exchanging data. Electronic Data Interchange (EDI) has also been used for exchanging data. However, EDI is rigid, complex, and expensive to implement. More recently some technologies have been proposed to allow a more object-oriented and less expensive approach, based on Remote Procedure Calls. Examples of these approaches are DCOM [12] and CORBA/IIOP [10]. The problems with these technologies are that they are platform specific, do not easily integrate, and pose network security risks due to the requirement of having open ports to accommodate messages.

The existing approaches have contributed to alleviate the problems of sharing data between autonomous and heterogeneous data sources. However, the development of Web services [17], SOAP [8] and XML technologies support the problems of e-business by allowing the ability of representing data structures and describing these structures in an easy way to implement and maintain. In the next section we describe an approach that uses Web services.

## 4. The approach

In order to tackle the problems described in Section 2 we proposed a lightweight approach based on Web services and related XML technologies. The

approach was developed in a way that can be implemented on multiple vendor platforms, with minimal effort and disruption to existing systems.

The main goal of the approach is to create an environment where new web-based information systems can be developed quickly and easily, using any technology platform, by accessing information from any of the existing 200 information systems at the FAO, and supporting the multilingual characteristics of the institution in which documents are expressed in five official languages as well as Russian and other local variations. Other objectives included the implementation of dynamic report generator and development of an XML document repository to handle metadata and language variants in a generic way.

In the next subsections we describe the *information bus* approach proposed to support data exchange and dynamic report generation.

### 4.1. Information bus

Figure 2 presents an overview of the architecture of the *information bus* being proposed to support interoperability of various information sources. The approach consists of wrapping the various data sources with Web service interfaces in which information inputs and outputs are passed as XML structures.

The concept of the *information bus* is that all data passed through it is represented in standard XML formats. These formats can be imposed in a regulated fashion by publishing the XML schemas being used and validating instances of messages. Regardless of the formats used by the existing systems, the same XML syntax is used for input and output parameters on the Web services. For example, all data related to country, language or currency is represented in a single XML format, which uses (a) ISO 3166 country code (3 letter), (b) ISO 639-1 language code (2 letter), (c) ISO 4217 currency code, respectively. With Web services it is not necessary to re-engineer existing systems to new XML standards. However, it is necessary to enforce XML standards in the Web services interfaces. For example, the parameters for operations involving language codes always use the 2-character ISO 639-1 code.

The Web services were developed for systems containing information about statistics, documents, maps, news and events. These systems are:

- internal to the FAO, for which the development team had access to the application source code,
- internal to the FAO, but the development team had no access to the application source code, and
- external to the FAO.

The management of information, including handling of multilingual variants is also based on XML. We propose to move structured information out of database fields and represent them in XML documents to allow a more generic model, which is easier to administer and to extend to new languages (e.g. there is a growing need to support Russian, in addi-

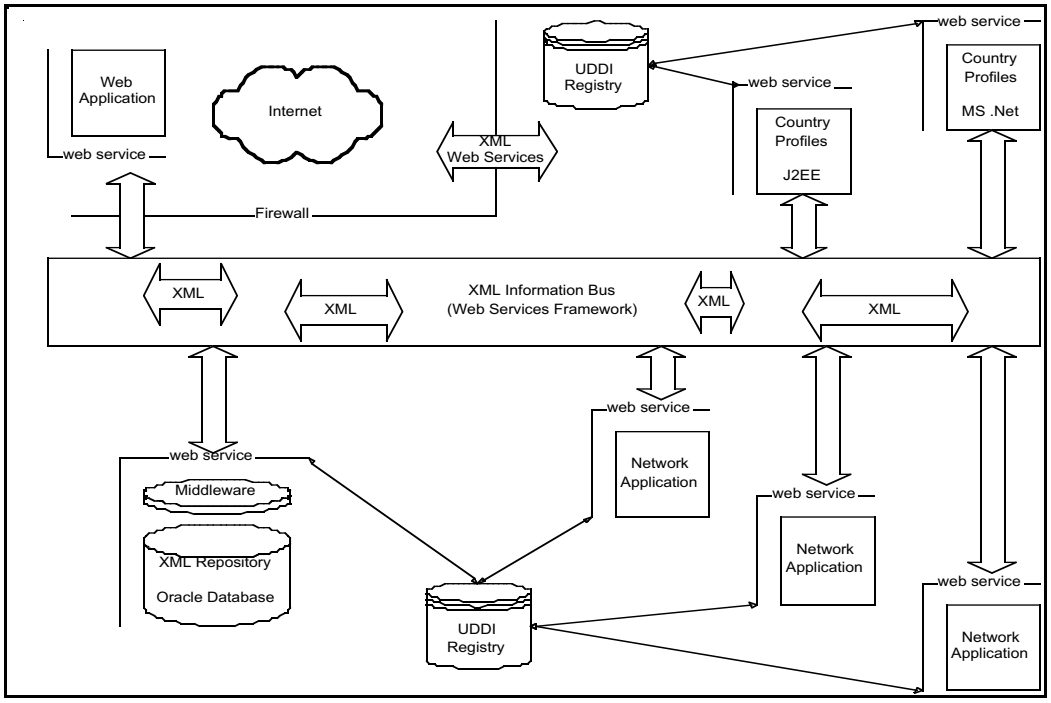



**Figure 2. Information bus architecture**

tion to the five existing official languages). Whereas existing systems use own (non-standard) database structures to model multilingual data, the XML approach provides a generic way to manage structured information to any schema. The XML documents are stored in an XML repository, as shown in Figure 2.

It is foreseen with the project proposal for the handling of metadata also to be based on XML and stored in the XML repository, which can be used as the exchange model throughout the FAO. The metadata would be represented as RDF [21], RDF Schema [7], Dublin Core elements version 1.1 [11], and XML Topic Maps [26]. RDF would be used to specify metadata on resources, i.e. values of properties for the resources. RDF Schema would be used to define classes of resources and the properties that instances of each class can take. In addition, RDF Schema, Dublin Core, and XML Topic Maps would be used to define ontologies, which capture the relationship between classes, resources, and properties that compose a vocabulary. XML Schemas [14] would also be used to define vocabulary range of values contained in a property.

The assignment of constraint metadata would be based on standard ontologies published or developed in-house, also represented in XML. This would facilitate importing and exporting of all XML metadata held in participating systems.

The XML repository stores resources (documents) in a relational database, using a Java interface based on an extended version of the XML:DB API [32] that caters for document variants (e.g. different language variants of the same document) and metadata associated with documents. The repository is also wrapped as a Web service to allow access of documents by metadata and/or language.

The FAO currently has a web application called FAO Country Profiles [15], which draws information from a variety of systems on the internal network and presents an aggregated view, sorted by country. Within each country profile, information is structured according to the main functional areas of the FAO - sustainable development, economic situation, agriculture sector, forestry sector, fishery sector, technical cooperation. We developed an application for the Country Profiles using the *information bus* architecture (see Section 5 for an example).

The architecture can contain two Universal Discovery, Description, and Integration (UDDI) registries to support discovery of information. One UDDI registry is internal to the FAO and assists with share and exchange of information between the data sources internal to the organization. The other UDDI registry is used to support share and exchange of data between the data sources external to FAO. In the initial deployment of the architecture, only the internal registry was active.

An example of the XML structure passed in the *information bus* is shown in Figure 3. It consists of a

```
<soap:Envelope
 xmlns:xsi="http://www.w3.org/2001/
                              XMLSchema-instance"
 xmlns:xsd="http://www.w3.org/2001/
                              XMLSchema"
xmlns:soap="http://schemas.xmlsoap.org/soap/
                               envelope/">
 <soap:Body>
  <Query xmlns="http://tempuri.org/">
     <Country>SEN</Country>
     <Language>EN</Language>
    <Keyword>Forestry</Keyword>
  </Query>
 </soap:Body>
</soap:Envelope>
```

**Figure 3. Example of XML structure passed in the information bus**

SOAP [8] message enriched with metadata from ontologies represented in RDF [21]. In this example, the XML structure represents a query about documents containing information of *forestry* (Keyword), in *Senegal* (Country – *SEN*), written in *English* (Language – *EN*). The transformation from the standard XML representation used in as the input parameters of the Web service, to the native input parameters of the system is implemented in the Web Service code itself. This is achieved using mapping structures from the native input parameters of the application (strings, integers) to the ISO representations outlined in the *information bus*.

In the approach, we propose to use three different types of Web services based on their functionality, named: *support*, *relevance*, and *content*. The support Web service type contains utilities to return standard representations of countries, metadata categories, and language translations. An example of the information returned by the support service is shown in Figure 4. The relevance Web service type is used to identify the Web service that is related to a particular application context and the setting of parameters necessary to call the identified Web service, as illustrated in Figure 5. In this example Web service with ID 900 contains description of general maps and should be accessed by using parameters such as *Country*, *Language*, and *Category*.

The content Web services type is invoked to return XML content from existing information sources, through Web services interfaces with parameters for language, country, subject, and others. Figure 6 shows an example of the content service returned from the BBC News Online information source (external to the FAO).

### 4.2. Country Profile report

Our approach also supports dynamic report generation based on data extracted from the various information sources. The reports are assembled as XML

```
<FSCollectionChoices
xmlns="http://tempuri.org/CollectionChoices.xsd">
 <Country diffgr:id="Country231"
          msdata:rowOrder="230">
   <COUNTRY>Zimbabwe</COUNTRY>
   <FS_COUNTRYCODE>181
          </FS_COUNTRYCODE>
   <ISOCODE>ZWE</ISOCODE>
 </Country>
 <Item diffgr:id="Item1"    msdata:rowOrder="0">
   <ITEM>Abaca (Manila Hemp)</ITEM>
   <FS_ITEMCODE>809</FS_ITEMCODE>
 </Item>
 <Element diffgr:id="Element1"
          msdata:rowOrder="0">
   <ELEMENT>Seed</ELEMENT>
   <FS_ELEMENTCODE>111
      </FS_ELEMENTCODE>
 </Element>
 <Year diffgr:id="Year1"   msdata:rowOrder="0"
       diffgr:hasChanges="inserted">
   <YEAR>1961</YEAR>
 </Year>
</FSCollectionChoices>
```

**Figure 4. Example of information returned from support service**

```
<ServiceDetails
 xmlns:xsi="http://www.w3.org/2001/
                    XMLSchema-instance"
 xmlns="http://tempuri.org/ServiceDetails.xsd">
 <ServiceDetail
     d2p1:ServiceName="GeneralMaps"
     d2p1:ServiceID="900"
xmlns:d2p1="http://tempuri.org/ServiceDetails.xsd">
   <ServiceDescription>
     Description for GeneralMaps
   </ServiceDescription>
   <Param d2p1:name="Country">
     <value>GBR</value>
   </Param>
   <Param d2p1:name="Language">
     <value>EN</value>
   </Param>
   <Param d2p1:name="Category">
     <value>16</value>
     <value>19</value>
   </Param>
 </ServiceDetail>
</ServiceDetails>
```

**Figure 5. Example of information returned from relevance service**

and rendered as PDF, by using XML Stylesheet Language: Formatting Objects – XSL:FO [1] and the open source FO Processor from Apache [4]. The reports are generated based on information content selected by the user.

When the user chooses a country and language from the support Web services this sets the state of the client and the relevance Web service is used to define the information available to the user in that context. Then when the user chooses to generate a dynamic report they are presented with the option to invoke different Web services, depending on the context. These Web services create the different sections of the report, according to the preference of the user.

Once the user has chosen the services to invoke in the creation of the report, the report generator calls all the Web services simultaneously using multi-threading. The report is built in memory in an order that depends on which Web service returns results first; the final report, in the correct order is compiled and generated once the last Web service returns results. The whole process takes approximately 60 seconds from invoking the services to report generation; a normal report will involve between 30 and 50 different Web services.

```
<BBCNewsDS
 xmlns ="http://www.fao.org/waicent/cpmis/
                              BBCNewsDS.xsd">
<BBCNews>
   <headline>Blair blasts green pacesetters
   </headline>
   <intro>In 1997 Labour undertook to be the
&amp;#34;first truly green government&amp;#34;,
but has that promise been fulfilled?</intro>
   <newsdate>23/10/2000</newsdate>
   <link>http://news.bbc.co.uk/hi/english/sci/tech/
         newsid_987000/987400.stm</link>
 </BBCNews>
 <BBCNews>
   <headline>Labour: A green government?
   </headline>
   <intro>In 1997 Labour undertook to be the
&amp;#34;first truly green government&amp;#34;,
but has that promise been fulfilled?</intro>
   <newsdate>23/10/2000</newsdate>
   <link>http://news.bbc.co.uk/hi/english/sci/tech/
         newsid_986000/986532.stm</link>
 </BBCNews>
</BBCNewsDS>
```

**Figure 6. Example of information returned from content service**

## 5. Example

In this section we present an example of the Country Profiles application [15] used as a case study for our approach. Country Profiles is an application in the FAO that allows access to country-specific information without the need to search individual databases and systems. It is an information retrieval tool that groups in a single area the vast amount of information available at FAO based on the global activities in agriculture and development, and classifies the information by country. The application uses three categories to group information:

1) FAO's areas of expertise - sustainable development, economy, agriculture, fisheries, forestry, and technical cooperation,
2) FAO's priority areas for interdisciplinary action (PAIA)– ranging from biological diversity to trade in agriculture, fisheries, and forestry, and
3) AGROVOC - a metadata ontology with over 4000 terms breaking down the first two metadata categories to a lower level (i.e. Cattle Breeding). AGROVOC is mainly used in the Library applications at FAO.

We have developed the Country Profiles application by using Web service technology. Figure 7 presents the web page used as the interface to the application. In Figure 7 it is possible to see all the different services used in the application.

Firstly the three dropdown lists under the banner at the top of the page are invoked from the support Web services described above. These set the state of the application and are currently set to English (EN), Afghanistan (AFG) and FAO's Fields of Expertise for the metadata. The categories to the left of the page (General Information) are also populated from the same metadata support Web service. Slightly below is the fourth dropdown list, this is populated using the relevance service, which takes inputs from the above three services and generates a list of available Web services which meet the current state of the application. It also contains the exact parameters to be sent to each content service when the user chooses an application (see Figure 6 for an example). Finally in the main body of the screen you can see an example of an invoked content Web service, in this example the service returns News information about the selected country from a system names EIMS.

Figure 8 shows examples of another content service being invoked (in this case the information in derived from the General Mapping application). Images are received in the body of the XML response as Base64 encoded string, which is decoded and cached on the client application server for faster retrieval. Legend text is also sent with the encoded string and the colors are generated using hash codes (e.g. #FFFFCC). The second screenshot in figure 8 shows a generated report in PDF format (see subsection 4.2).

## 6. Implementation aspects and evaluation

An operational prototype tool has been implemented in the period of three months in order to demonstrate and evaluate the approach. The case study used in the development was the Country Profiles application, as illustrated in Section 5. The prototype allows (a) generic XML-based information infrastructure to support multilingual information in an extensible and standard way, (b) application integration structure based on Web services to allow interoperability of FAO systems and information sources for delivery through web portals, (c) use of Microsoft .NET [25], (d) standard XML representations for handling metadata and multilingual documents, and (e) dynamic country profiles report generation. In addition, the prototype has also demonstrated the ability to combine information in multiple languages together in the same pages, to develop new web-

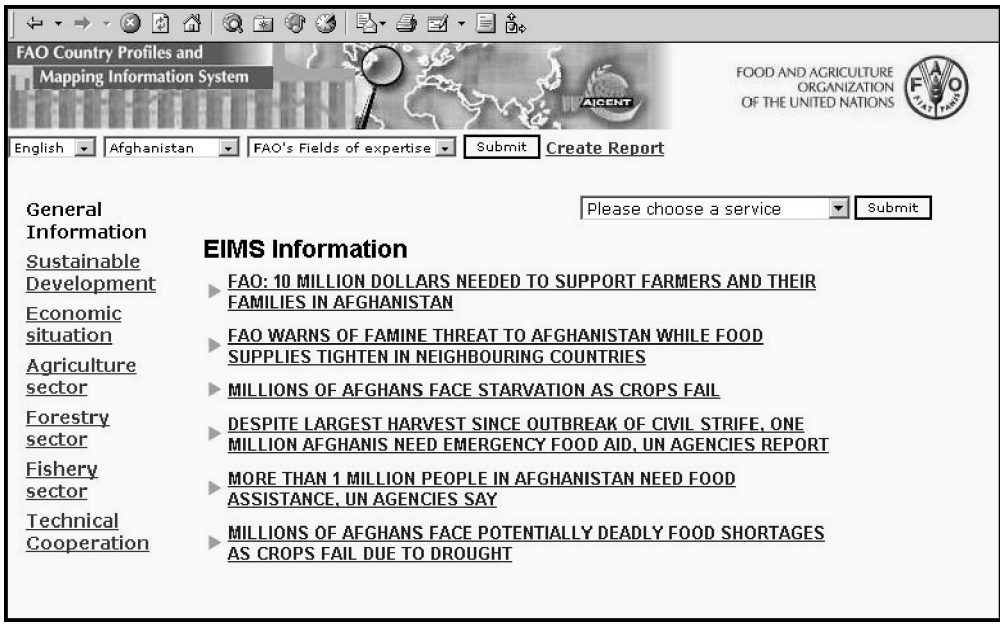



**Figure 7. Example of the web page for the Country Profile application**

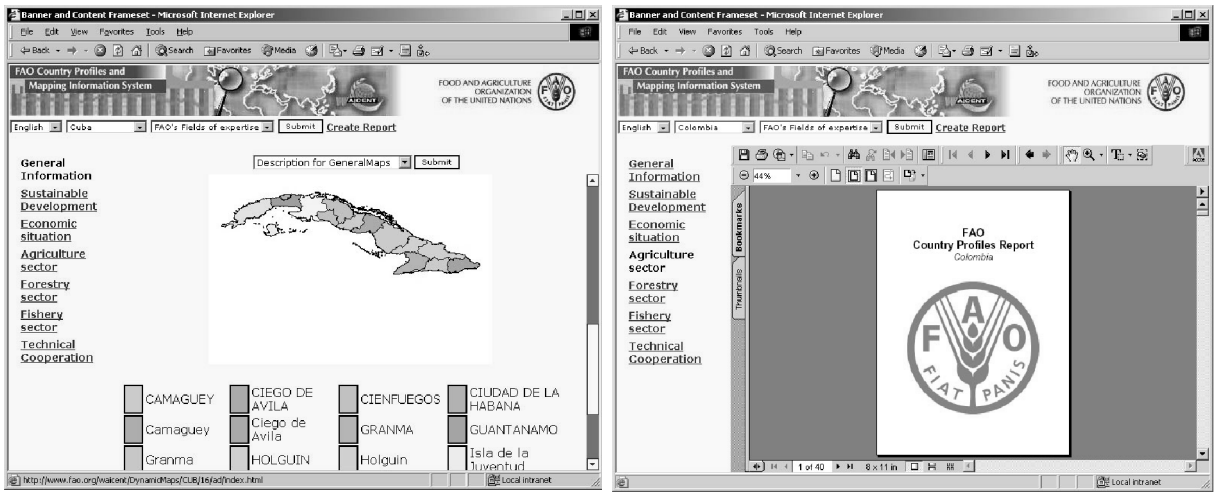



**Figure 8. Screenshots of the Country Profile application**

based information systems quickly and easily, and to combine and compare statistics from different countries.

The main activities in the development of the prototype consist of the (a) creation of Web services wrappers to existing data sources so that they could be accessed by using the XML *information bus*, (b) implementation of a new XML document repository that allows structured data to be stored for different languages in a generic and extendible manner, and (c) implementation of the Country Profiles application.

The Web services wrappers were created using Microsoft .NET for the following systems (those already accessible from the web are listed with their URL):

- Statistics
  a) FAOSTAT - an internal FAO statistics system (http://apps.fao.org) and
  b) World Bank Statistics – an external system
- Documents:
  a) FAOBib - an existing internal FAO bibliography system (http://www4.fao.org/faobib)
- EIMS (Electronic Information Management System) - an existing internal FAO repository of full text documents
- RAP - a new internal XML document and metadata repository
- Maps
  a) General Maps – an internal FAO application of maps
  b) GeoNetwork – an internal FAO geographic application (http://www.fao.org/geonetwork)
- News and Events
  a) NEMS (News and Events Management System) - an internal FAO news application
  b) BBC News Online - an external new service

For EIMS, RAP, and NEMS systems the development team had access to the applications source code and respective databases. For FAOBib, GeoNetwork, General Maps, and BBC News the development team did not have access to the source code and the information was accessed by HTTP on scrape HTML data. For FAOSTAT and World Bank Statistics the access to the data was through batch and cache.

Our experience has been very positive. We have found that it was easy to develop the wrappers around the data sources. Some of the activities have been implemented in hours, instead of days, as it was previously thought.

A major advantage of using Microsoft .NET framework was the ease with which Web services wrappers could by created. However, the integration of these Web services with the J2EE platform had some problems due to the difference in handling complex data types and inconsistencies in the use of Web Services Description Language (WSDL).

One problem is related to the fact that .NET uses Document-style Web services by default, whereas the J2EE implementation (Apache Axis) uses RPC-style invocation. To alleviate this problem in .NET we used the SoapRpcService() [30] property to indicate that the .NET Web service was RPC-style. However, there were further problems because Axis did not yet implement support for multi-dimensional arrays or for generating complex type definitions in WSDL, which were created automatically by .NET. To alleviate these problems, and to allow developers to create Web services quickly and easily from existing Microsoft applications (of great importance to FAO, to encourage all departments to make their applications available as Web services), a second tier of Web services was created that automatically made the transformation from the data types generated by .NET to XML arrays that could be used by both .NET or J2EE Web services.

The prototype has shown that it is possible to integrate different information sources (internal and

external to the organisation) by preserving their autonomy and that the system can evolve in an easy way by adding and removing data sources. In addition, it has also demonstrated that it is possible to avoid the problem of imposing the use of a single technology in an organization like the FAO. The Web services framework used in our approach allows a platform that is stable, flexible, extensible, and high performance.

The work presented in this paper has provided new opportunities for the FAO. Examples of these opportunities include, but are not limited to, standardization on the way information is shared within FAO and with external parties, provision of new information services within FAO (e.g. ontologies, statistics presentations), provision of an environment to efficiently develop, deploy, and maintain new information services, leading of a next generation information dissemination methods to assist with the aims of the FAO.

## 7. Conclusion and future work

In this paper we have presented an *information bus* approach to interoperate different data sources in the Food and Agriculture Organization of the United Nations. The data sources are both internal and external to the FAO and are used to help developing countries to modernize and expand agriculture, forestry and fisheries by collecting, analyzing and disseminating information which can be used to fight hunger and achieve food security. The different data sources contain information related to different types of documents written in five official languages, statistical data, electronic bibliography references, maps and graphics, and news and events.

The approach is lightweight and based on Web services and XML technologies. It preserves the autonomy of the existing systems and allows evolution by adding and removing data sources. The approach allows the creation of an environment where new web-based information systems can be developed quickly and easily, supports the multilingual characteristics of an institution like FAO, provides dynamic report generation, and handles metadata and language variants in a generic way.

Before development of a full implementation of the prototype throughout the FAO, we are extending the prototype to support XML configuration files, generic report configuration tailored by subject area, ontology services in which based on a data item a list of related data is identified, and public Web services for ontologies, countries, and codes. We are investigating the application of Dublin Core to represent metadata information and, therefore, extending its current use on supporting metadata associated with documents in the XML repository. We also plan to implement the Country Profiles application by using J2EE Web services technology and compare this technology with Microsoft .NET.

## Acknowledgement

We would like to thank all members of the WAICENT/FAOINFO Dissemination Management Branch team at the Food and Agriculture Organization of the United Nations. Giorgio Lanzarone, who acted as the project coordinator in FAO, and Marta Iglesias, Nick Waltham and Anne Aubert for participating in the requirements specification phase and developing the Web services wrappers in the data sources of FAO. We would also like to thank some members of CSW Informatics. Mavis Courname for participating in the requirements specification phase and Stephen Horwood responsible for the implementation of the country profiles applications, Web services client wrappers, and XML repository.

## References

[1] Adler, S., Berglund, A., Caruso, J., et al., 2001. Extensible Stylesheet language, Version 1.0, W3C Recommendation, October 2001, http://www.w3.org/TR/2001/REC-xsl-20011015.

[2] AGROVOC. http://www.fao.org/agrovoc.

[3] Ahmed, R., Albert, J., Du, W., Kent, W., Litwin, W. and Shan, M.C., 1993. An Overview of Pegasus. *In the 3rd International Workshop on Research Issues and Data Engineering: Interoperability in Multidatabase Systems,* pages 273-277, Vienna, Austria, April 1993. IEEE Computer Society Press.

[4] Apache. Apache Project. http://www.apache.org.

[5] ASP. Microsoft Active Server Pages. http://www.microsoft.com/asp.

[6] Bray, T., Paoli, J., Sperberg-McQueen, C.M. and Maler, E., 2000. Extensible Markup Language (XML) 1.0, Second Edition, W3C Recommendation, October 2000. http://www.w3.org/TR/2000/REC-xml-20001006.

[7] Brickley, D. and Guha, R.V., 2002. Resource Description Framework (RDF) Schema Specification 1.0, March 2002. http://www.w3.org/TR/rdf-schema.

[8] Box, D., Ehnebuske, D., Kakivaya, G., Layman, A., Mendelsohn, N., Nielsen, H., Thatte, S. and Winer, D. Simple Object Access Protocol (SOAP) 1.1. http://www.w3.org/TR/SOAP.

[9] Chung, C., 1990. DATAPLEX: An Access to Heterogeneous Distributed Databases. *Communications of the ACM*, 33(1): 70-80, January 1990.

[10] CORBA/IIOP. Common Object Request Broker Architecture. http://www.omg.org/technology/documents/formal/corba-iiop.htm.

[11] Dublin Core. http://dublincore.org.

[12] DCOM. Microsoft Distributed Component Object Model,
http://www.microsoft.com/ com/tech/DCOM.

[13] Dwyer, P.A. and Larson, J.A., 1987. Some Experiences with a Distributed Database Testbed System. In *Proceedings of the IEEE*, volume 75, pages 633-648, May 1987.

[14] Fallside, D.C., 2000. XML Schema Part 0: Primer. http://www.w3.org/TR/2000/WD-xmlschema-0-20000407.

[15] FAO. FAO Country Profiles and Mapping Information System.
http://www.fao.org/countryprofiles/.

[16] Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassalos, V. and Widom, J., 1995. The TSIMMIS Approach to Mediation: Data Models and Languages. In *Next Generation Information Technologies and Systems (NGITS)*, Naharia, Israel, June 1995.

[17] Graham, S., Simeonov, S., Boubez, T., Davis, D., Daniels, G., Nakamura, Y. and Neyama, R., 2002. Building Web Services with Java: Making Sense of XML, SOAP, WSDL, and UDDI. SAMS Publishing, 2002.

[18] Grimson, J. and Murphy, J., 1995. The Jupiter Approach to Interoperability with Healthcare Legacy Systems. In R.A. Greenes, H.E. Peterson, and D.J. Protti, editors, *MEDINFO 95*, pages 367-371, IMIA 1995.

[19] Heimbigner, D. and McLeod, D., 1985. A Federated Architecture for Information Management. ACM Transaction on Office Information Systems, 3(3): 253-278, July 1985.

[20] Java. Java JSP/Servlet. http://java.sun.com.

[21] Lassila, O. and Swick, R.R., 1999. Resource Description Framework (RDF) Model and Syntax Specification. February, 1999.
http://www.w3.org/TR/REC-rdf-syntax.

[22] Litwin, W., 1988. From Database Systems to Multidatabase Systems: Why and How. In W.A. Gray, editor, *In Proceedings of the 6th British National Conference on Databases (BNCOD 6)*, British Computer Society Workshop Series, pages 161-188, July 1988.

[23] Litwin, W., Mark, L. and Roussopoulos, N., 1990. Interoperability of Multiple Autonomous Databases. ACM Computing Surveys, 22(3): 267-293, September 1990.

[24] Liu, L. and Pu, C., 1997. Dynamic Query Processing in DIOM. Bulletin of Technical Committee of Data Engineering, 20(3): 30-37, September 1997.

[25] .Net. Micosoft ASP .Net. http://www.asp.net.

[26] Pepper, S. and Moore, G. XML Topic Maps (XTM) 1.0. http://www.topicmaps.org/xtm/1.0.

[27] Roth, M.T. and Schwatz, P., 1997. Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources. *In Proceedings of the 23rd International Conference on Very Large Database Bases*, Athens, Greece, August 1997.

[28] Sheth, A.P. and Larson, J.A., 1990. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Surveys*, 22(3): 183-236, September 1990.

[29] Smith, J.M., Bernstein, P.A., Dayal, U., Goodman, N., Landers, T., Lin, K.W.T. and Wong, E, 1981. Multibase – Integrating Heterogeneous Distributed Database Systems. In *National Computer Conference*, volume 50 of *AFIPS Conference Proceedings*, pages 487-499, 1981.

[30] SOAP-RPC. http://www.soaprpc.com.

[31] Tomasic, A., Raschid, L. and Valdueriz, P., 1996. Scaling Heterogeneous Databases and The Design of Disco. In *Proceedings of the 16th International Conference on Distributed Computing Systems (ICDCS)*, pages 449-457, Hog Kong, May 1996. IEEE Computer Society Press.

[32] XML:DB. http://www.xmldb.org.

[33] Zisman, A. and Kramer, J., 1999. An Approach to Interoperation between Autonomous Database Systems. *Distributed Systems Engineering Journal*, 6, 1999.

[34] Zisman, A., 1998. Information Discovery for Interoperable Autonomous Database Systems. PhD Thesis, Department of Computing, Imperial College of Science Technology and Medicine, University of London, UK, 1998
(http://www.soi.city.ac.uk/~zisman).

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 157-162*


# Design of a Federation Service for Digital Libraries: the Case of Historical Archives in the PORTA EUROPA Portal (PEP) Pilot Project

Marco Pirri, Maria Chiara Pettenati, Dino Giuli
m.pirri@telemat.det.unifi.it, pettenati@achille.det.unifi.it, giuli@det.unifi.it
Electronics and Telecommunications Department
University of Florence
Via Santa Marta 3, 50139 Florence (IT)

## Abstract

*Access to distributed and heterogeneous Internet resources is coming up as one of the major problem for future development of the next generation of Digital Libraries. Available data sources vary in terms of data representation and access interfaces, therefore a system for federating heterogeneous resources accessible via the Web is considered to be a crucial aspect in digital libraries research and development. Libraries as well as institutions and enterprises are struggling to find solutions that can offer the final user an easy and automatic way to rapidly find relevant needed resources among heterogeneous ones.*

*Our project starts from the recent results of Dublin Core Metadata Initiative (DCMI) and in particular from the Dublin Core Recommendations (DCMIR).*

*This paper reports our analysis of three different digital historical archives maintained by the European University Institute (EUI) in Florence and its mapping using a common Meta Resource Card based on Dublin Core Elements (DCMES). This situation requires careful consideration of interoperability issues related to uniform naming, metadata formats, document models and access protocols for the different data sources.*

*We also present our Porta Europa Portal (PEP) federated architecture that will support an XML Dublin Core implementation and in our aim should be easily open to RDF future support. The PEP pilot project specialised portal should provide high quality information, selected according to the criteria of originality, accuracy, credibility together with the cultural and political pluralism derived from the EUI's profile. The information in Porta Europa will be relevant, reliable, searchable and retrievable.*

**Keywords:** *Federated service, digital libraries, Dublin Core, Metadata, interoperability*.

## 1. Introduction

The integration of existing digital libraries and electronic catalogues of publication is considered to be one of the major issues for the digital library community. The purpose of digital library integration is to devise a proper architecture, a metadata structure and a suitable protocol to:

- provide a uniform interface hiding the specific features and restrictions of the single sources;
- supply integrated view on the data.

These issues are tied to two main aspects (Endig et al. 2000):

1) the access to data sources (the digital library) depends on the query interface and capabilities of specific data source which have therefore to be carefully described;
2) a specific data format is used in each single digital library, therefore mapping into a common format is required.

State of the art in digital libraries has shown an evolution of data integration approach along two main directions (Hanani & Frank 2000): from the Stand-alone Digital Libraries to Federated Digital Libraries. In the first case the Digital Library is maintained by a single institution and the data collection is self-contained while the material is localised and centralised. The second case is related to a federation of several independent Digital Libraries in the network, possibly organised around a common theme or topic. The Federated Digital Library regroups many autonomous Stand-alone Digital Libraries forming a networked library accessible through a unique user interface.

The digital library federation service approach is therefore adopted to cope with this issues of data integration where the need of regrouping different

Stand-alone Digital Libraries arises such it is the case of this project. It is worth remarking that, even if the archives are managed by a single institution, such it is our case, the digital libraries are considered to be stand-alone because of their heterogeneity in metadata, document formats and access interfaces as it will be more clearly explained in the sequel.

The interoperability issue is consequently decomposed in the sub-problems related to uniform naming, metadata formats, document models and access protocols.

This paper reports on the preliminary study for the design of a federation services for the integration of three different digital libraries (here also referred as *data sources*) – three heterogeneous archives related to historical topics – whose access has to be made uniform through a single portal: the Porta Europa Portal.

## 2. The History Pilot Project - The Porta Europa Portal

The PEP (Porta Europa Portal) Pilot Project refers to the integration of three digital libraries related to European history topics: Voices on Europe, Virtual Library and Biblio library catalogue.

Each of these data source is characterized by:

- A collection of data objects (digitized audio, html pages, records ...) available locally or through the network
- A collection of metadata structures
- A collection of services (access methods, management functions, logging/statistics, etc.)
- A domain focus (topic)
- A community of users

The need of integrating the three data sources comes from the topic (European history) and users community which are common to all three archives.

- **Voices on Europe;** (http://wwwarc.iue.it/webpub/Welcome.html) Voices on Europe is an archive containing the electronic audio version and electronic transcriptions about a hundred of interviews given by outstanding politician and historians.
- **WWW-VL (Virtual Library) on European History Integration;** (http://vlib.iue.it/history/index.html) The Virtual Library (VL) is the Web oldest catalogue, conceived by Tim Berners-Lee. Unlike commercial catalogues, it is run by a loose confederation of volunteers, who compile pages of relevant links for specific areas in which they are expert. The EUI Library Web site contains the complete list of VLs belonging to the **WWW VL History Project** in the University of Lawrence/Kansas (USA) and mirrored at the European University Institute's Library (EUI).
- **Biblio (the EUI historical archives);** (http://www.iue.it/LIB/Catalogue/) This is the library catalogue containing more than 250.000

**Table 1. Main properties of the three data sources**

| Characteristics of the archive | Voices on Europe | Virtual Library | Biblio Library Catalogue |
|---|---|---|---|
| Data objects | Digitized audio-video tapes<br>Interviews written transcription (pdf) | HTML pages | Records |
| Collection of metadata structures | The archive is organised in Access Database | The archive is structured in Web pages | The archive is maintained in a proprietary database in USMARC format |
| Collection of services | The access to the interviews is currently performed via a Web interface through SQL queries.<br>Resource management is allowed directly on the database.<br>No logging or statistic functions are allowed. | The access is performed through the Web, maintenance ad updating of the information is managed through the Web by a project administrator. No logging or statistic functions are allowed | Information management functions are performed through INNOPAC Library automation system. |
| Domain focus (topic) | European history | | |
| Community of users | Everybody for information search<br>On a case basis, restricted access for full documents consultation<br>Administrators for information management | | |

bibliographic records. Access to resources is supported by INNOPAC, well known Library Automation System (INNOPAC).

As it is remarkable by the properties illustrated in Table 1, the heterogeneity of the three data sources are due to their difference in the types of data objects, in the collection of metadata structures and in the collection of services provided by each access interface. It is therefore clearly outstanding the need to provide a federation system to integrate access and management of the archives.

## 3. The PEP Project development phases

The cultural and operational context of the European University Institute and the presence of a top class library in the social sciences with an emphasis on European issues brought to the idea of building a **specific Portal Project** integrated inside the EUI Web Site and offering opportunities to link the currently dispersed European oriented information sources and to contribute also to a better visibility of the Institute. The proposal is to create a specialized portal - **Porta Europa** - which should answer to this need and position the Institute itself on the Web as a leader in the "European debate" and as a natural gateway, a logical point of access to high quality information on European issues.

To test the feasibility and the impact of the PEP project the EUI committed itself to the development of a PEP prototype concerning historic topics. To this extent, among the various available digital historical archives three of them were chosen for the implementation of the pilot, as described in the previous paragraph.

The PEP Pilot Project is being developed according to the following steps:

*1. Analysis of the three data resources*

In this part we analysed the current situation of the resources and we identified the main issues involved in each case. Each resource is characterised by different issues which have been elicited and therefore faced (see Table 1). This phase ended with a detailed description of the metadata formats, document models and access protocols for each of the data sources. The analysis revealed the strong points and the weakness of each digital library setting the basis for the definition of a common document description model. More specifically we defined a Meta Resource Card (MRC) with a detailed mapping of the relevant fields derived by each resource. Table 2 illustrates a synthesis of the MRC where each archive single fields are more detailed in the related internal reports to be shortly published by the EUI library (Pirri and Noiret 2002) (Pirri and Terzuoli 2002) (Pirri and Baglioni 2002).

*2. Definition of the federation architecture*

After the first phase, the analysis and definition of the federation architecture has to be covered. According to what available in literature (Endig et al. 2000) we agreed on the conceptually layered architecture described in paragraph 4, where each layer has to provide/use specific operation to/from adjacent layers. The objective of the federation services architecture is to provide uniform interface to the individual resources and to supply an integrated view on the data. Therefore the architecture must be conceived in order to accept queries on the global view

**Table 2. Resources Mapping in Meta Resource Card**

| Dublin Core Element | Voices on Europe | Virtual Library | Biblio |
|---|---|---|---|
| Title | Interviewee's surname/name | Title | Title |
| Creator | Name of Interviewer | Author | Author |
| Subject | Level 1,2,3 (eurovoc) | Type 3 | Subject |
| Description | Full text Interview | Abstract | Note |
| Publisher | Eui | Type 1 | Imprint |
| Contributor | Not used | Not used | Not used |
| Date | Date of recording | Date of insertion | Date of pubblication |
| Type | Video/Audio/Testo | Text (Html) | Text |
| Format | Pdf | Html | Pdf |
| Identifier | Url | Url | Isbn |
| Source | Not used | Not used | Not used |
| Language | Language | English | Lang |
| Relation | Additional Material | Not used | Not used |
| Coverage | Not used | Not used | Not used |
| Rights | User Profile | Free | User Profile |

**Table 3. Users and related roles for information access**

| Function | User |
|---|---|
| General Administration (information management) | Administrator and Project Leaders |
| Information search | Public |
| Full information access | Internal users (IUE member, professors, students, etc.) |
| Restricted information access (restriction is due to property right on some resources contained in the archives) | External users, groups of users |
| Personalised services | Registered users |

(uniform data model), decompose them and translate them to allow processing from the single data sources.

*3. Definition of the user roles for information access*

Due to the variety of information accessible through the different digital libraries of this project, an important step consists in the definition of the users role and access rights.

For the scope of this project, we can identify the functions reported in Table 3 which are to be associated to the related users in order to allow the maximum flexibility in the management and access to the resources.

Users functions and roles have been used in the archives analysis phase as for the Dublin Core Rights field and will also be used in the next development of the project.

## 4. The PEP federated architecture

The architecture of our federation service (Endig 2000) is structured in three layers: the *data source layer* where all information is stored with autonomy of representation and access interfaces, the *adapter layer* were special adapters (harvesters) have to be implemented to provide uniform access and transform the data source specific model into the global model of the federated system, and the *federation layer* which is responsible for global data integration using an on purpose database.

**Data Source Layer**: these are the archives (digital libraries) whose integration we deal with: Voices on Europe, Virtual Library and Biblio library catalogue.

**Adapter Layer:** this layer provides uniform access to the information, hiding the differences in the data models and query interfaces. Here the metadata are mapped from the source specific model into the global model of the federated system - the Meta Resouce Card derived according to the Dublin Core Elements.

Relevant work has been done in literature as for the role of the Adapter layer. At this stage of the project we are considering the possibility to use the approach defined in the Open Archive Initiative (OAI) (Lagoze & Van den Sompel 2001) (Lynch 2001) where the Data Sources function as a Data Provider adopting OAI technical framework to expose metadata about their content. On the other side Service Providers (for instance the Federation Service) harvest metadata from data providers using the OAI protocol, to provide value-added services. According to this approach the Adapter layer would implement all the Harvester and the OAI protocol.

It is worth highlighting that the OAI approach addresses the interoperability issues requiring that all data providers (Data Source) provide the metadata in a common format, namely the Dublin Core Metadata Element Set (Weibel 1998). This approach has been adopted in successful initiative concerning digital libraries federation (Liu et al. 2001).

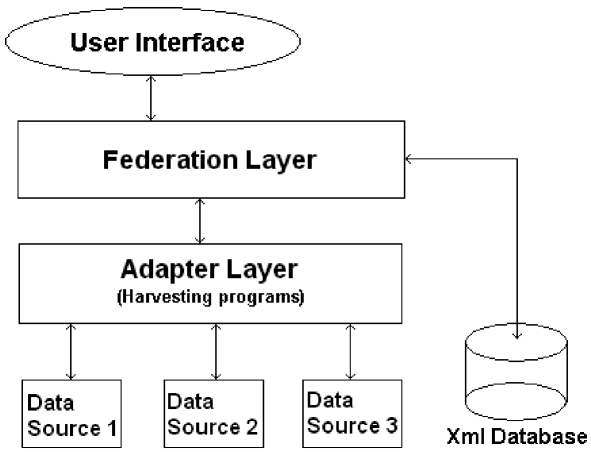



**Figure 1.**

**Federation Layer:** in this layer the services for definition and query of the integrated data vision are provided. Metadata describing information of the three different resources are stored in a unique XML database.

To this extent a common metadata format (Meta Resource Card - MRC) must be devised for the three resources. To effectively address the interoperability issue, the Meta Resource Card should follow the unqualified Dublin Core Standard to define the common fields. This choice is compliant with the Open Archive Initiative intentions.

We are also investigating the possibility to find Federation layer solutions capable to become easily compliant with RDF approach.

On top of the Federation Layer we added the User Interface which will provide information access through the Web to all the users. The use of active pages will allow service personalization, according to the user's role and the actual function exploited as reported in Table 3.

## 5. Conclusion

This paper reports on the design of a federation service for three heterogeneous digital libraries. The scope of the federation service is to provide a common metadata format for gathering information from the available data sources and to provide a unique querying interface to access them.

At this stage of the project we analysed the state of the art in order to choose the most suitable realisation approach accounting for sound theoretic issues such as Dublin Core Metadata and Open Archive Initiative which are now being investigated in the digital libraries community. Our purpose is also to devise a simple yet easily realisable solution to validate the pilot requirements.

The three data sources analysis is now completed, highlighting the major differences of the three archives.

We therefore choose a federated model with a consequent layered architecture aiming at implementing the OAI protocol and the Dublin Core Metadata description.

We defined a Meta Resource Card, according to the Dublin Core Standard, to unify the description of the federated data to the PEP user.

We are now continuing the realisation of the pilot project whose the first results are expected by autumn 2002.

## References

DCMI, Dublin Core Metadata Initiative, OCLC, Dublin Ohio.
http://dublincore.org/

DCMIR, Dublin Core Metadata Initiative Recommendationst
http://dublincore.org/documents/

DCMES, 1999. Dublin Core Metadata Element Set, Version 1.1: Reference Description
http://dublincore.org/documents/dces/

Endig M., Hoding M., Saake G., Sattler K.U. and Schallehn E., 2000. Federation services for heterogeneous digital libraries accessing cooperative and non-cooperative sources. *In:* International Conference on Digital Libraries: Research and Practice, 2000 Kyoto, 120 -127.

Hanani U., Frank A.J., 2000. The parallel evolution of search engines and digital libraries: their convergence to the Mega-Portal. *In:* International Conference on Digital Libraries: Research and Practice, 2000 Kyoto, 211-218.

Lagoze C., Van de Sompel H., 2001. The Open Archives Initiative: Building a low-barrier interoperability framework. *In:* the ACM/IEEE Joint Conference on Digtial Libraries, Roanoke VA June 24-28, 2001, 54-62.

INNOPAC Official Web site http://www.iii.com/

INNOPAC Users Mailing List http://innopacusers.org/

Liu X., Maly K. Zubair M., Nelson M.L., 2001. Arc - An OAI Service Provider for Cross Archiving Searching. *In:* the ACM/IEEE Joint Conference on Digtial Libraries, Roanoke VA June 24-28, 2001, 65-66.

Liu X., Maly K., Zubair M., Nelson M.L., 2001. Arc - An OAI Service Provider for Digital Library Federation. D-Lib Magazine 7(4).

Lynch C., 2001. Metadata Harvesting and the Open Archives Initiative. ARL Monthly Report 217, August 2001, http://www.arl.org/newsltr/217/mhp.html.

OAI The Open Archives Initiative Protocol for Metadata Harvesting. The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14.
http://www.openarchives.org/OAI/openarchivesprotocol.html

Pirri M., Noiret S., 2002. WWW VL EUI History Project Report and Analysis for future development in the PORTA EUROPA Portal (PEP) Pilot Project.

Pirri M., Terzuoli G., 2002. Voices on Europe for PORTA EUROPA Portal (PEP) Pilot Project.

Pirri M., Baglioni P., 2002. Library Automation System Summary on INNOPAC Manual Description for PORTA EUROPA Portal (PEP) Pilot Project.

Rdf, Resource Description Framework (RDF)
http://www.w3.org/RDF/

Weibel S., 1998. The Dublin Core: A simple content description format for electronic resources. NFAIS Netwsletter, 1998, 40(7), 117-119.

Xml, Extensible Markup Language (XML)
http://www.w3.org/XML/

Paper Session 4

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 165-169*


# Describing Services for a Metadata-driven Portal

John Roberts
Archives New Zealand
NZGLS Custodian
john.roberts@archives.govt.nz

## Abstract

*This paper describes New Zealand E-government activities supporting the discovery of services through the use of Dublin Core-based New Zealand Government Locator Service (NZGLS) metadata. It notes the issues faced in collecting service metadata from agencies to populate a new whole-of-government portal. The paper then considers the adequacy of the metadata schema for service description, and identifies a difficulty in applying definitions which refer to the content of the resource to a process-like resource such as a service. Three approaches to this challenge are suggested: creating a surrogate description to provide a source of content; treating the information exchanged in conducting the service as the content; and using additional contextual metadata. The adequacy of the schema for covering all the users' needs for discovering and using a service is examined, and the need for metadata about specific service delivery points and conditions is noted. Finally, it is observed that future stages of e-government will require more sophisticated descriptions of services to support processes beyond discovery.*

## 1. Introduction

In a paper to the DC-2001 Dublin Core conference[1], the treatment of services was identified as a critical area for the use of discovery metadata by the New Zealand E-government programme. Discussions in the DC-Government Working Group confirmed that E-government programmes worldwide are seeking to take a service-centric approach to representing government to the public. This paper explores the importance of service description for e-government, and a range of issues in applying Dublin Core-based metadata in this way. It draws on experiences of the New Zealand E-government programme in using the Dublin Core-based New Zealand Government Locator Service (NZGLS) schema in the development of a new whole-of-government portal which is both service-focused, and metadata-driven.

## 2. Background – NZGLS

Recognising the need for standardised metadata to support resource discovery across the whole of government, New Zealand Government officials developed the NZGLS discovery level metadata standard based on the Dublin Core Metadata Element Set and the Australian Government's extension of DC, the Australian Government Locator Service (AGLS). NZGLS includes the same four additional elements by which AGLS extends DC: function, availability, audience and mandate. Along with AGLS, the versions of NZGLS released to date have provided explicit guidance on the application to services. However, this is one of the least stable areas of the standard, as there is as yet no clear consensus on what precisely we mean by "a service", or how the public articulate their discovery needs for these resources.

In December 2001 the New Zealand Government formally adopted this local adaptation of DC as "the official New Zealand Government standard for creating discovery level metadata in the public service". The Cabinet decision directed public service departments to make NZGLS compliant metadata available "to ensure that *their services* and relevant information resources (both online and offline) can be discovered by the Portal search engine's metadata searching capability" (emphasis added). A new portal is scheduled for public launch in August 2002, and relies heavily on NZGLS service-metadata. A companion paper explores in more detail the practical experiences involved in the development and implementation of the Portal and the related metadata management facility[2].

## 3. Services in E-government

In the words of the UN Report *Benchmarking E-government: A Global Perspective*[3], "Services are the public face of government". That report is typical of e-government literature in linking the success of e-gov-

ernment with ICT-based service discovery and delivery. The report recognised the New Zealand Government's achievements in this area by ranking New Zealand third behind the US and Australia in its global E-Government index. However, no jurisdiction was ranked at the top stage of e-government development, "seamless" (described as total integration of e-functions and services across administrative and departmental boundaries). Significantly, the higher levels of e-government maturity are generally described as demonstrating an integration of service delivery with technology, building on technology-based tools for service location, discovery and description.

## 4. E-Services Project

To address the range of issues around services as part of effective e-government and in the first instance to gather structured information about services delivered by New Zealand public service agencies, an early project initiated by the New Zealand E-government Unit was its e-Services Project[4]. Phase One of this project was to compile an inventory of services, by getting all agencies to describe the services they provide. Emphasis was placed on the description of services from a customer perspective, as the information has been used initially for delivery of a new whole-of-government portal. Later phases look for additional opportunities to move service delivery on-line and to integrate delivery channels.

The service descriptions were produced using the NZGLS metadata standard, though the details of the elements were obscured from the users, and masked by more service-focused terms and a plain-language interface.

In part because of this business focus, the project targeted business analysts and communications staff to create the metadata records, rather than web masters or librarians. 55 agencies were included in the initial collection, and around a thousand services were identified and described. An assumption was that in the first instance, agencies are in the best position to identify and define their services.

Among the issues to emerge were:

- Consistency of service "size": some agencies identified a large number of low-level services, more akin to interactions, while others identified a small number of very broad services, more like functions.
- Consistency of description: where comparable services are delivered by several agencies, how can you ensure consistent descriptions. This is of particular importance in cases where the same responsibility is exercised by different organisations in different parts of the country.
- Multi-agency services: how to develop a single descriptive record for a service which, as thought of by the public end-user, comprises actions and decisions of multiple agencies working together.

The management of these issues is considered in more detail in the companion paper.

## 5. Services in Dublin Core

The DCMES is clear that it is intended to be applicable to descriptions of services: "For the purposes of Dublin Core metadata, a resource will typically be an information or service resource". Similarly, "Service" is a defined term in the DCMI-Type vocabulary, defined as "a system that provides one or more functions of value to the end-user. Examples include: a photocopying service, an authentication service, interlibrary loans, a Z39.50 or Web server". However, the use of the DCMES for true service description appears under-developed compared with its high profile for information resource discovery purposes. For example, no specific guidance on the use of the DCMES with service-type resources is included in the *Usage Guide*.

## 6. Describing Services

With services taking a central role for e-government, and practical experience through exercises such as the e-Services Project, the opportunity now presents for reflection on how well NZGLS (and DC) has handled the description of services, and to consider whether the issues that have emerged relate to the metadata model itself (core definitions and semantics), to the tool used to facilitate the collection and management of the metadata records, or to the training support and expertise of those creating the records.

What, then, actually are services? First we should be clear that we are not talking about services in the same sense that web service description initiatives (eg Web Service Description language, WSDL[5]) use the term. Services are ongoing, and they have an activity dimension to them, they represent ways of doing business. Services are by definition transformational – it is the provision of something of value to a user that is the essence of a service. The New Zealand e-Services Project defined a service as something that "provides value (tangible, experienced, or information) to a service user. The service may be provided directly or through a contracted supplier, and can be delivered via one or more transactions". Elsewhere, similar definitions have been used in other jurisdictions. In Australia, the Commonwealth Government considers "a service exists where a relationship is established between a business function of a government agency and the identified needs of an individual or group. Examples of government services are family allowance assistance, grants programs and the receipt of payments by government agencies such as the Australian Tax Office. The AGLS metadata obligations in the Government Online

Strategy require agencies to describe all services, regardless of mode of delivery"[6]. The Queensland State Government defines services as "the activities undertaken by an agency on a repetitive basis either to fulfil legislative requirements or to satisfy an external client need".[7]

The provision of information (a service) is qualitatively different from the piece of information provided (an information resource). Services are abstract, and exist in the eye of the beholder. Yet services are meaningful, indeed they represent a view of the world more directly related to the way citizens think about their dealings with government than models focused on organisational structure or information resources. The New Zealand experience suggests that the description of services will be an important aspect of e-government, but that our current understanding of how to apply metadata standards to this class of resources is not sufficiently robust to avoid risks of inconsistency.

In light of this, is it possible and useful to describe services directly and explicitly? There is a subtle difference between describing the service itself, and describing a document about a service. This distinction can, however, be critical. Think of the case where a web page outlines a service: the identifier or language would relate to the document, and are meaningless to locate or understand the service itself. The first issue to address in service description is clarity around the resource that is being described: is it the service itself (however defined) or a (related) document? This distinction is easily blurred when the documents themselves contain the information which will also feature in a metadata record for the service (eg a web page describing a service and its availability), or where the document is a necessary part of using the service (eg an application form). Similarly the distinction between the service and the service provider (ie the agency delivering the service) can be easily confused. Only when these distinctions are well understood can the best application of the elements and the possible refinement of their definitions be evaluated.

User feedback strongly indicates that public users like an interface which groups resources into topic clusters, an expectation perhaps shaped by experience with commercial directory-structured portals and search sites. It would be ideal to construct such groupings automatically from *Subject*, *Function*, and other elements of our metadata. These groupings should effectively be saved searches, built from queries that reflect the logic that users would employ in their discovery paths. In practice this has not proved to be fully possible for us on the basis of the metadata provided by agencies alone. This reflects in part the challenge of achieving sufficient consistency in a devolved metadata authoring environment. To ensure the discovery interface supports user needs, these groupings have been directly included in the *Subject* element from a controlled set of "Portal groups". That this has been necessary suggests that our existing metadata doesn't (yet) adequately reflect the goal of "what the user thinks the service is about". This aim may be made more difficult by inconsistent approaches to such questions by our users.

One of the issues in applying the canonical DCMES definitions is the concept of the content of the resource. Seven of the element definitions (*Creator, Subject, Description, Contributor, Type, Language, Coverage*) relate to the content of the resource, yet the very idea of the content of a service is problematic. Consider, for example, *Subject*. Which information should one consider to determine the subject of a service resource? As the service model is a user-driven representation, perhaps the most useful way of approaching this is to consider what the public user would think the service is about. In practice, this equates to describing the service, and using that description as the content-object which is analysed for a subject. A description of how the service is delivered and what it delivers to the user (the value the user obtains, to go back to our definition) would act as a surrogate piece of content. For the *Description* element at least this could become quite circular!

The above approach has limitations, and an alternative starts from the defining characteristic of services in providing value through interactions. Putting these interactions centre-stage suggests we should consider the information that flows between participants when the service is used. The content referred to in element definitions is then the content of the exchange, that is, the information handled by the service process. For some elements, this approach is appealing. In the case of *Language*, for example, the potential user wants to know in what language(s) they will be able to interact with the service – it is the language of the flow that is important. The subject then would be the subject of the conversations represented by the conduct of the service.

A third approach is to supplement the content analysis and content-oriented elements with contextual information. Rather than seeking a way of finding sufficient meaningful intrinsic content for description, this would focus on where the service delivery occurs within the extrinsic context environment of government activities and structures. This approach is based on the principle that the key for users is how to obtain service delivery. Service discovery then could be considered primarily about identification of delivery channels and points, rather than "content". This approach is likely to require additional elements. Building on the *Creator* and (particularly) *Publisher* elements, which relate the content to its environment, additional contextual elements such as *Function*, *Availability* and *Mandate* attempt to deal with the limitations of a content analysis model for services. *Function* relates the service to the reason for providing that service.

These ideas relate mainly to the application of existing definitions to service descriptions. The approaches should not be seen as competing models, from which one should be selected. Rather, the question is where each should be used in order to meet user needs. The above discussion relates to activity in the domain of the metadata creator. But discovery metadata is primarily about meeting the needs of the end user. Considering the matter from the other side then, how adequate is the framework for representing the important characteristics of services?

For the citizen, the critical factors are the ability to locate information about the service that meets the need they have, and to determine how they can access the service (where, when, at what cost). This goes beyond a narrow sense of discovery (confirming the existence of a resource) to a view of discovery that includes the full chain of search and evaluation behaviour through to use. Several of the DCMES elements relate to characteristics of an information resource which will enable a potential user to assess their ability to access and then use the resource. *Format* allows a judgement to be made in terms of the ability to extract meaningful content from a particular manifestation or medium, including hardware and software dependencies. *Publisher* covers the person or organisation which makes the resource available, but it appears to stretch the semantics of this element to use it to also cover details of how and where a service resource is delivered. *Format* can include the channels through which a service is delivered, but is better fitted to dealing with this in the general sense (eg delivery by counter service, by free phone etc) rather than as a means of setting out specific delivery points and service hours. NZGLS and AGLS include a specific *Availability* element to address this question.

These complexities highlight the importance of the relationships between resources, and of recognising the richness of those relationships. Simple metadata models capture flat representations of reality. Services are less easily forced into such models than information resources. The boundaries of individual services along a service chain may be difficult to distinguish. From different perspectives a wide range of different articulations of the extent of a given service (just which transactions comprise a single describable bundle) may be valid. These considerations make the *Relation* element crucial to the development of robust service description approaches. Similarly, this element provides space for linking service descriptions with metadata records for other types of entity (eg linking the service metadata to descriptions of necessary forms or procedural documents). The ongoing discussion about the representation of Agent details in Dublin Core metadata is relevant in this space. Sophisticated use of *Relation* enables considerable complexity to be represented within a comparatively simple metadata schema. Its potential to help address the issues of service description is still largely unrealised.

## 7. Relating the Service View to other Representations

The service view is only one way of describing government to the public. As well as traditional organisational models of government, more recently we have seen an emphasis in many jurisdictions on output-based models. The output model is similar in many ways to a service model, but is typically linked to financial accountability structures in an inward-looking way that sits uncomfortably with the more fully outward, citizen-focused view represented by the service model.

Description in service terms not only provides a tool for the public to understand and interact with government, but provides an additional tool for government itself to analyse its activity. Service descriptions created to aid discovery will be used to support other forms of evaluation. Is there duplication or overlap of services? Which services may be priority candidates for e-enablement? How can the model help with organisational design considerations? These questions will come into sharper focus as exercises like the New Zealand E-services project move into their later phases, and as governments move through the UN's e-government development levels. Many services share common underpinning business processes, such as registration, payment, or application lodgement. How do business processes intersect with service description and discovery? It is probable that metadata resources created initially for discovery purposes will then be challenged and pushed into service to support other aspects of the e-government agenda. It remains unclear what further service metadata elements may be needed in these future stages of e-government maturity. The New Zealand E-services project, for example, collected transaction volume data as part of its information gathering, but the long term value of this and other characteristics of services remains to be seen.

## 8. Evaluation of Public Response

A new New Zealand Government portal will be formally launched in August 2002, release having been delayed by the announcement of a general election which was held in late July. By the time of the Dublin Core Metadata Conference in October, it is expected that there will be preliminary user feedback, and evidence from web logs to enable some informed comment to be made about the public response to the portal, and about the usefulness of service metadata as used and presented in the portal.

## 9. Conclusion

Discovery metadata is inherently sensitive to the perspective of the end-user. Refinement to our meta-

data models may be required to provide the details need by users who are seeking to discover services rather than information resources. There is value in providing an integrated discovery framework for a range of resource types, including both services and documents. What is needed is refinement rather than comprehensive change, however the extent and nature of that refinement is as yet still open for discussion. A range of tactics exists for incorporating service metadata in existing models. These all show potential for sharpening our understanding of service description, and for addressing different aspects of the challenge of effectively representing services in a readily discoverable and meaningful manner. The approaches discussed in this paper are not mutually exclusive – the question is rather which to use where. Experience gained through implementation exercises such as the New Zealand Government portal will help inform and steer these developments.

---

[1] John Roberts, *Between a Rock and a Hard Place: Dealing With NZGLS Development Issues*, in *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001*, (2001) <www.nii.ac.jp/dc2001/proceedings/product/paper-42.pdf>.

[2] Sara Barham, *New Zealand Government Implementation of a DC-based Standard – Lessons Learned, Future Issues*.

[3] United Nations, Division for Public Economics and Public Administration, *Benchmarking E-Government: A Global Perspective*, (2002) <www.unpan.org/e-government/Benchmarking%20E-gov%202001.pdf>.

[4] Further information at <http://www.e-government.govt.nz/e-services/index.asp>.

[5] The W3C Web services Description Working Group defines a web service "a software application identified by a URI [IETF RFC 2396], whose interfaces and binding are capable of being defined, described and discovered by XML artifacts and supports direct interactions with other software applications using XML based messages via internet-based protocols". Further information at <http://www.w3.org/2002/ws/desc/>.

[6] *Commonwealth Implementation Manual: Australian Government Locator Service (AGLS) Metadata* <http://www.naa.gov.au/recordkeeping/gov_online/agls/cim/cim_introduction.html#1.5>.

[7] <http://www.iie.qld.gov.au/comminfo/download/is34.pdf>.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 171-176*


# New Zealand Government Implementation of a DC-based Standard – Lessons Learned, Future Issues

Sara Barham
Portal Information Manager
E-government Unit, State Services Commission
Sara.barham@ssc.govt.nz

## Abstract

*This paper summarises key implementation issues encountered with the New Zealand Government's discovery level Dublin Core-based metadata standard, NZGLS. In particular, it discusses the processes used to create and manage NZGLS-compliant metadata throughout New Zealand's core public service agencies. This metadata is being used to support the New Zealand government's new service-focussed portal.*

## 1. Introduction

This paper covers the implementation of the New Zealand Government Locator Service (NZGLS)[1] into the New Zealand public sector, as part of the development of a new whole-of-government portal[2]. A companion paper notes the issues faced in collecting service metadata from agencies[3].

## 2. New Zealand Government Portal Strategy

In 2001 the E-government Unit[4] of the State Services Commission completed a Government Portal strategy. It outlined a vision for a portal which would (1) give people and businesses access to information and services provided by the web sites of individual government organisations, and (2) include guidance about how to find information and services that are not available via the Internet. To succeed, the portal will direct people to government information that is always current and accurate. This means government organisations must keep their web sites and metadata current and accurate.

## 3. Use of Metadata

The need for high quality, consistent descriptions of services and documents became paramount as a result of this strategy. The obvious consequence was the development and implementation of a government metadata standard, the New Zealand Government Locator Service, based on Dublin Core and the Australian Government Locator Service (AGLS).

## 4. Government Agency Commitment to the Strategy

The most critical success factor in the achievement of this portal strategy has been a commitment from government agencies to develop, own and manage their metadata records. The principle of agency ownership of their service and document descriptions is fundamental to the ongoing success of the portal. In order to achieve a well-populated portal within the bounds of available time and funding the E-government Unit has worked intensively with a "critical mass" of agencies. This includes about 45 central government agencies, 5 quasi-government agencies (such as Accident Compensation Corporation) and 10 major local government jurisdictions (such as Auckland City Council). Interestingly, as a result of work with the "critical mass", the metadata from about 30 other closely related agencies has also been included on the portal in the initial implementation. It is intended that other agencies will add their metadata over time.

## 5. Achievement of the Goal

A combination of "push and pull" (otherwise known as require and encourage) strategies, driven by the E-government Unit, has led to an exceptional level of commitment from the agencies to deliver high quality metadata about their services.

Specifically this has included agencies:

- Attending two rounds of training in both service description and use of a metadata creation tool;

- Creating over 1500 service descriptions and 2000 document descriptions;
- Developing an understanding of the NZGLS standard;
- Developing internal mechanisms and processes to develop a service-based representation of their work; and
- Committing to long-term internal management processes for their metadata.

### 5.1 Architecture

At the heart of the Metadata Management Facility (Metalogue) is the Portal Metadata Repository. This repository is also the key link, or interface, with the E-government Portal. Authoring of the metadata takes place in the centralised metadata repository, but is devolved to agencies. The metadata repository is required to store, and manage, NZGLS metadata elements. A user-initiated search of the portal involves searching two sources of data (1) the metadata repository and (2) an index of all New Zealand government web sites. The MMF is also integrated with two customised New Zealand Government thesauri, Subjects of New Zealand (SONZ) and Functions of New Zealand (FONZ). The next stage of development of the MMF includes more "workflow" components such as easier access for an agency to an overview of its own records and their status in the flow from authoring to Portal; a communication space for agency metadata creators and agency userid management.

### 5.2 Training

About 250 people from government agencies have participated so far. The first training course was about the definition of services. The second course (on the use of *Metalogue*, the Metadata Management Facility) was developed in two phases following a training needs analysis. First, a strategy was developed, in consultation with the E-government Unit, by SWIM Ltd[5]. This strategy was then used by another consultancy firm, The Sysdoc Group Ltd[6], as the basis for writing a course to deliver to agency representatives. To date, all training costs have been met by the E-government Unit to ensure that appropriate momentum is achieved in the creation of metadata.

### 5.3 Creation of service and resource descriptions

Beginning in October 2001 agencies were introduced to the concept of "E-services". This process assisted agencies to list and describe their services from a client perspective. Until April 2002 agencies worked on these descriptions using a customized service description wizard utility, called the Services List Tool Set. This Tool Set was also developed by the Sysdoc group Ltd. At this point, the use of metadata elements was not introduced directly. Service analysis and description was the main focus. However, in almost all cases elements used to describe services had an equivalent NZGLS element. When these service description records were migrated from the Tool Set to *Metalogue*, the new Metadata Management Facility, the NZGLS elements were prepopulated from these equivalent fields. For example, one Tool Set element was named Agency Name; its equivalent from the standard is Creator.

By the time the data was migrated, there were about 1000 service descriptions from a core group of government agencies. With the advent of Metalogue in late April 2002, agencies were then able to refine their service descriptions using, for example, the controlled value lists for elements such as Subject. Agencies were greatly assisted by the fact that the MMF now enforced the NZGLS standard to a much greater extent.

### 5.4 Understanding the NZGLS standard

Knowledge of the standard across the government sector in October 2001 was, at best, patchy. A number of representatives of government agencies had assisted in the development of the standard, and they and some of their colleagues knew what the standard was, and how it would be used. These people tended to be based in agencies' information management groups. But the vast majority of agency people did not have that same understanding. For example, there were many representatives from agency communication groups and business units. In the E-government Unit's experience, the word "metadata" was bound to either cause terminal boredom to set in very quickly, or to panic otherwise calm and resourceful people!

Awareness of the NZGLS standard was developed in the following ways: referring to it during the Service Listing process, including providing an electronic link to the standard, but using "real" language to communicate metadata concepts; more detailed training in its use during the Metadata Management Facility training, constant reference to the standard as the basis of metadata compliance in communication with agencies and via a Cabinet mandate for its use. By early 2002 the term "metadata" had been used several times by the Minister of State Services in public speaking engagements or Cabinet meetings!

### 5.5 Using the standard in "real life"

There is an inevitable gap between the standard on paper, and how it is interpreted and used. The aim of the E-government Unit and the Custodian of the standard (Archives New Zealand) is to take a pragmatic approach to its use. A good example of the distinction between NZGLS, the implementation of NZGLS in *Metalogue* (the Metadata Management Facility) and the use of the metadata by the Portal is the following: an issue which has challenged all con-

cerned is the creation of separate records for documents in different formats, such as PDF, HTML or hard copy. In operational terms, this has required double or triple the effort from an agency to create records which adhere to the standard. The current effect of this on the display of Portal search results is to show several related information resources with exactly the same titles. The E-government Unit and the NZGLS Working Group are examining the options for a pragmatic solution to this, while still remaining true to the standard.

A similar issue which has challenged us is in relation to use of the Date element. There has been confusion around which encoding scheme to use for this element, the difficulty partly being caused by the way it has been set up in Metalogue, where both ISO 8601 and DCMI Period appear in the same dialog box.

#### 5.6 Developing internal processes to ensure accurate representation of agency services on the portal

Agencies have had to develop new ways of working internally to ensure that their services are represented appropriately and accurately. Coordination between information management, information technology, communication groups and business units within an agency has been achieved in a number of different ways. One of the main principles being followed by agencies is that business units take responsibility for agreeing to and releasing service descriptions to the Portal. We believe most agencies understand that the multidisciplinary aspect of this process challenges existing ways of working, of integrating their Web presence into core processes.

#### 5.7 Committing to long-term internal management processes for agency metadata

One area yet to be tested is the ability of agencies to manage their metadata long-term. Already, for example, we have seen agency website reengineering causing broken links to appear in service records. The E-government Unit is committed to maintaining high quality information on the portal and therefore, a broken links report is being run regularly. But the main point is agencies having the same commitment to metadata maintenance and management. They need to put processes in place which ensure that any changes to services, either content or access-related, are also reflected in the metadata.

## 6. Quality Assurance Process to date

The metadata collection process has involved two phases of record creation and a major centralised quality assurance process. From late April to mid-June, agencies created additional service records and the majority of their document records. From May to July agencies received detailed feedback on the quality (both content and achievement of NZGLS standard) of their records from staff at the E-government Unit. This was a labour-intensive process, with a team of 6 fulltime staff working directly with agencies. The E-government Unit was committed to this approach to ensure that the metadata was of a high and consistent standard which could be redeveloped and amended when necessary, with a high level of confidence in the integrity of the data. This collaborative process has ensured that, for the most part, agency ownership of the metadata records has been achieved.

## 7. A Portal based on Metadata

The New Zealand Government portal relies heavily on metadata for its searching and its content. A subset of the 19 NZGLS metadata elements is focused on by the portal search mechanism (for example, Title, Description, Subject, Function, Rights, Relation.Requires, Availability and Audience). The content of many of these elements forms the content of the portal's search results, but also, significantly, points users to an agency's own website where more detailed information such as contact details for physical offices, specific application forms or brochures, are available. A topic hierarchy approach to locating information leads searchers by category to the information they want. If they reach a "dead end", that is, they do not find what they want, the following message appears: "Have you found what you want? If not search for x". An automatic search will then be generated, based on the topic name.

## 8. Benefits and Outcomes of the Strategy

Some of the unanticipated results of this implementation process are cross-agency communication and cooperation; higher visibility to policy and operational agencies of all government activities; strong commitment to the portal from local government agencies and agencies discovering for the first time where other agencies encroach in their operational areas. Agencies have gained much from opportunities to share their experience and solve issues together.

We expect to establish a Metadata Management support network to enable agencies to continue to have these opportunities. The portal provides a view of government services from a cross-agency perspective; gaps in and duplication of services will become more obvious. Local government (regional, city or town councils) has taken the opportunity to develop a centralised profile for its services, for citizens who make no distinction between local and central government service availability.

## 9. Key Reasons for Success

- A "Trojan horse" approach to metadata – agencies' initial introduction to the NZGLS metadata elements was masked within a service description wizard utility; agencies became familiar with the concept of describing their services using consistent rules, without necessarily realising they were following a metadata standard. By the time the standard-driven MMF was introduced, agency representatives had become more comfortable with the notion of metadata.
- Dedicated support for agencies via a Metadata Collection project team at the E-government Unit (training, helpdesk, documentation, feedback on quality of metadata records); a tailored training course, deskside assistance and a first-level helpdesk provided agencies with expertise, encouragement and support as they created their own records. Notably, the first two phases of the services listing and metadata collection took about three times the predicted effort by the E-government Unit in spite of a high degree of commitment by agency staff at all levels.
- Intensive work done with local government representatives to (1) produce a list of generic services, that is, services provided by all local authorities and (2) create agreed titles and descriptions; local government agencies were brought together centrally to coordinate a response. For example, one of the most popular services on the Portal is "Find location of public toilets around New Zealand". These are facilities provided by all local government authorities, and can be described effectively as a collective service, rather than as eighty or so individual services.
- A metadata capture tool which, to a large extent, enforces the NZGLS standard; and
- A Cabinet mandate for the use of the NZGLS standard; in December 2001 the New Zealand Cabinet "*agreed* that use of the NZGLS Metadata Standard be the official New Zealand Government standard for creating discovery level metadata in the public sector" and "*directed* all Public Service departments … to become NZGLS compliant (as specified in paragraph A.1.1 of the NZGLS Metadata Standard), and make NZGLS metadata records available to the NZGLS System [Metalogue], so as to ensure that the services and relevant information resources (both online and offline) can be discovered by the Portal search engine's metadata searching capability".

## 10. Some Future Issues

1. Conducting a reality check-review results of portal usage; assess impact on metadata creation, implement changes;
2. Using already created metadata elements, e.g. Coverage, to produce a regional, user-centric view of government services;
3. Maintaining the momentum to acquire metadata from additional agencies and broadening the coverage of metadata from existing "enrolled" agencies;
4. Managing the relationship between broad-based use of NZGLS and more detailed sector focussed metadata;
5. Transfer of metadata between the MMF repository and agencies for other purposes; and
6. Creating a balance between agency and centralised maintenance of portal metadata – whose metadata is it, anyway?

## 11. Conclusion

Producing the metadata building blocks for a new New Zealand Government portal is a significant achievement. The coordinated approach across government agencies is ensuring commitment to joint ownership of the portal. An upcoming challenge is to translate user response to the portal's structure and content into manageable changes to the now-considerable body of existing standard-based metadata.

---

[1] The New Zealand Government Locator Service (NZGLS) Metadata Standard and Reference Manual http://www.e-government.govt.nz/nzgls/standard/index.asp

[2] New Zealand Government Portal http://www.govt.nz

3 John Roberts , *Describing Services for a Metadata-driven Portal*, Paper presented to DC2002

[4] New Zealand E-government website http://www.e-government.govt.nz

[5] SWIM Ltd http://www.swim.co.nz/

[6] The Sysdoc group Ltd http://www.sysdoc.co.nz/

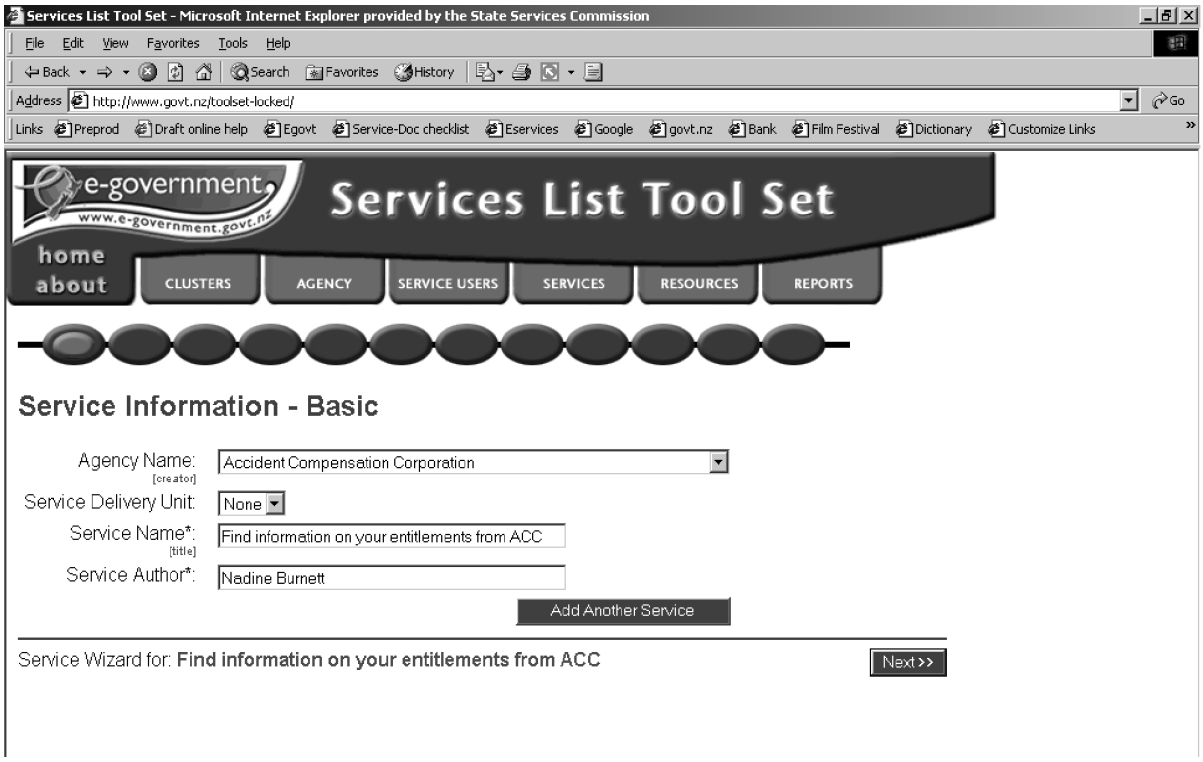



**Figure 1. Services List tool set screen**

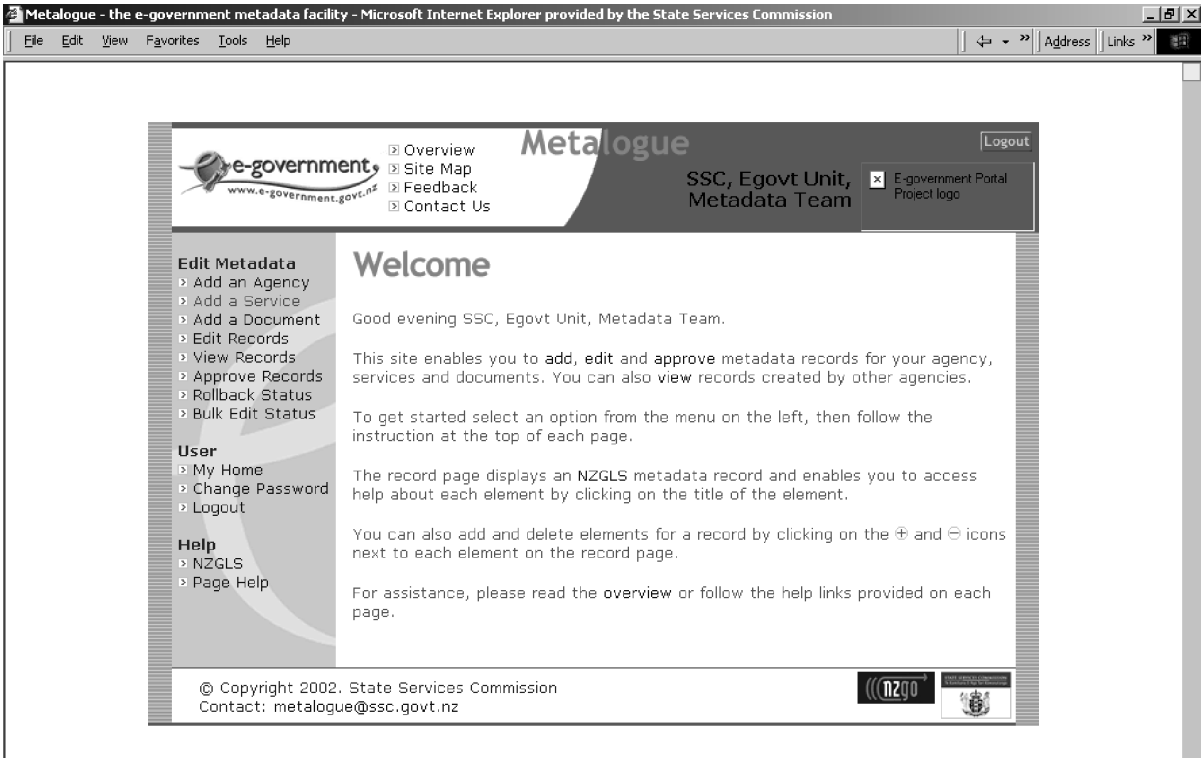



**Figure 2. Metalogue Welcome Screen**

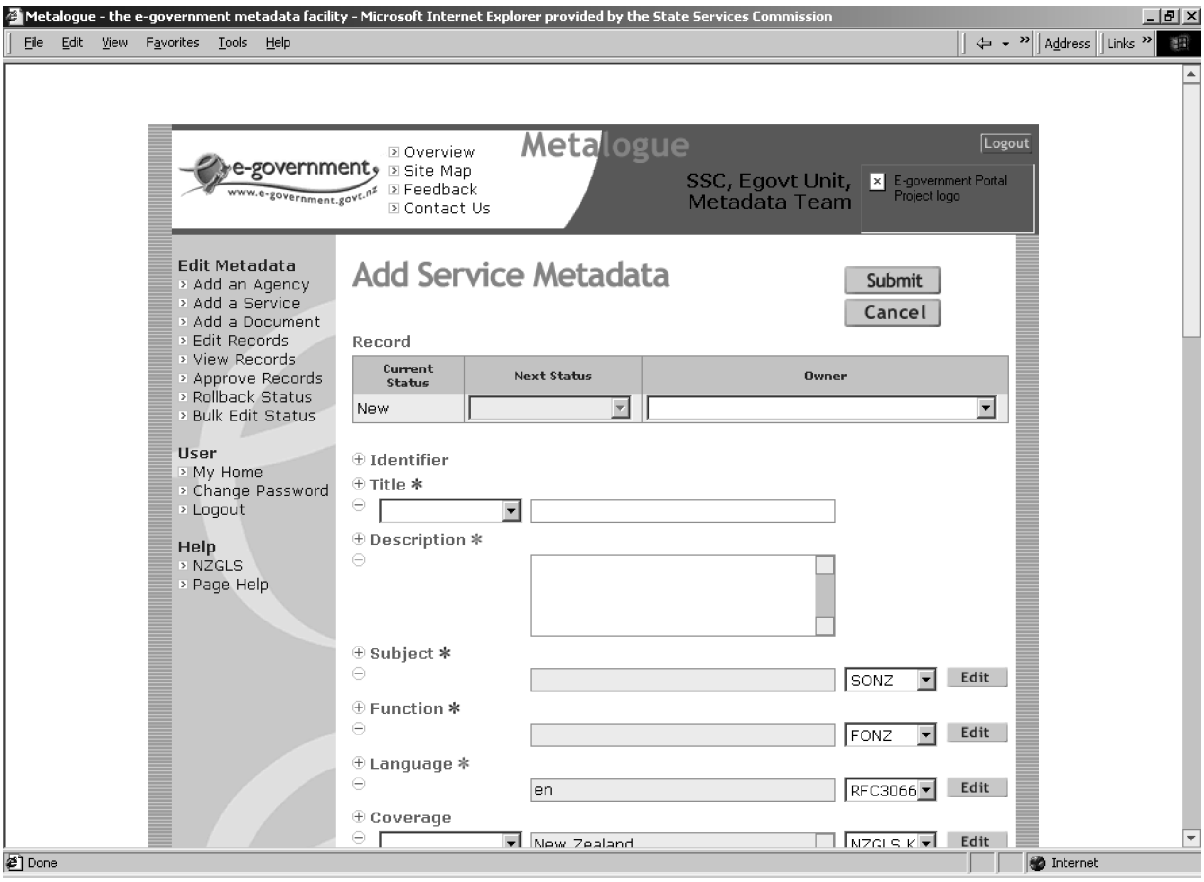



**Figure 3. Metalogue Add Service screen**

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 177-183*


# Visualising Interoperability: ARH, Aggregation, Rationalisation and Harmonisation

Michael Currie, Meigan Geileskey, Liddy Nevile, Richard Woodman
mfcurrie@ozemail.com.au, Meigan.Geileskey@dpc.vic.gov.au,
liddy@motile.net, richard.woodman@dpc.vic.gov.au

## Abstract

*This paper proposes a visualisation of interoperability to assist real-world deployment of metadata.*

*For some time, resource managers in many organisations have been acting on faith, creating 'standards compliant' metadata with the aim of exposing their resources to provide interoperability in discovery activities. In some cases, their faith has led them to miss the very essence of the work they are doing, and they have not got what they worked for.*

*The authors report a case study involving government agencies in Victoria, Australia. A number of departmental agencies have implemented, more or less, the DC-based Australian Government Locator Service (AGLS) application profile, at least for their web resources. They have done this with care and precision, with the long-term aim of developing a fully interoperable system. In the case study, typical would-be records for seven government departments were studied and it was shown that the tiniest, and typical, variation in use of the standard can be expected to thwart the aims of interoperability in significant ways.*

*In the context of the government's move to seeking interoperable metadata for all resources, including those within document management systems, the authors make visible how a small 'creep' can lead away from interoperability and how it might be contained in the future. They use a 3-step approach of 'aggregation, rationalisation and harmonisation' to expose the problems with 'nearly good enough' interoperability and the benefits of good interoperability, and encourage true harmonisation.*

**Keywords:** *Metadata, interoperability, aggregation, harmonisation, rationalisation, Dublin Core, government, AGLS.*

## 1. Introduction

This paper aims to describe a method used to demonstrate how even small variations in the interpretation and use of standards can affect interoperability efforts. It proposes a visualisation of interoperability, particularly in order to make it more understandable to less-expert metadata managers. The managers in the authors' context were information managers of government departments, and the aim was to develop metadata that would lead to the discovery of each and every document, or resource, in a government intranet as might be required if a minister were questioned in parliament on a particular topic. The aim of the project was to encourage the managers to strive for increased interoperability. It shows how easily the move to local applicability among the various agencies has led away from interoperability in practice and how it might be contained in the future. To do this, the authors use a 3-step approach of 'aggregation, rationalisation and harmonisation' and expose problems with 'nearly good enough' interoperability.

## 2. Literature review

While there is a lot of literature on the mechanics of interoperability, there is very little that attempts to make it understandable to non-specialists, or connect it with everyday practice..

From the very start, the ability to search across a range of resources was a high priority. A number of authors have attempted to define and explain the function of interoperability but they generally do so in the context of metadata that is not fully interoperable. They are working on strategies for bringing together collections. Typically, Hunter (2001) states that interoperability is intended to "enable a single search interface across heterogeneous metadata descriptions, to enable the integration or merging of descriptions which are based on complementary but possibly overlapping metadata schemas or standards and to enable different views of the one underlying and complete metadata description, depending on the user's particular interest, perspective or requirements". Arms (2002, p. 3) argues that "the goal of interoperability is to build coherent services for

users, from components that are technically different and managed by different organizations".

In the case of a single author, a government, working through its many agencies within an intranet, the differences between the collections can be expected to depend upon the different domains of operation. This does not necessarily mean they should be technically different even though variation may occur in levels of granularity of description of content.

The 1995 minimalist standard known as the Dublin Core was designed to meet the basic needs of different communities for "specifying metadata to support cross-domain resource discovery on the Internet" (Weibel, 2000). The Dublin Core Metadata Element Schema (DCMES) operates within the extensive Warwick Framework (Lagoze, 1996) which provides a modular structure to DC metadata to enable it to accept not only metadata from other standards but local qualifiers to existing elements and separate elements to meet the specific needs of the client groups. In order to maintain consistency in adding qualifiers, interoperability strategies involve carefully defined structure and registries to record and make available the different local applications. The composite view of DCMES is endorsed by what are now known as 'application profiles' (Heery, 2000). While the use of application profiles provides a solution to ensuring local specificity, current use of these often limits interoperability rather than enhances it, as shown below.

Heery (2000) argues that there is often disparity between the practices of the standards makers and the implementers. The former group views the power of metadata in consistent adherence to the accepted standards. Implementers, on the other hand, need metadata that serves their specific needs. Seeking to produce an effective, differentiated service, they often assume that the restrictions of adhering to accepted standards limit the utility of metadata to their users. The authors investigated this issue and propose that by sticking to standards, in the case of an intranet, the agencies involved could achieve both the power and the utility they seek.

Hunter (2001) argues that "significant new initiatives ... are demanding application profiles which combine elements from a number of different existing standardized metadata schemas whilst maintaining interoperability and satisfying their own specific requirements through refinements, extensions and additions" (p. 1).

Problems of interoperability can lie in meeting the needs of different standards. In highlighting the need to reconcile the competing/complementary needs of DCMES and INDECS/DOI, the rights focussed metadata adopted by the publishing community, Bearman, Rust, Weibel, Miller and Trant (1999) proposed using a common logical model, the IFLA Functional Requirements for the Bibliographic Record (FRBR). "Translating both the INDECS requirements and the DC requirements into the IFLA model provided the framework of a common logical expression for the two perspectives (in which) common semantics can be identified for each metadata element" (p. 6). This approach, using a third model to promote the interoperability of two others, has been replicated in a number of contexts.

In another approach, Blanchi and Petrone (2001) propose yet another digital architecture for managing and sharing metadata and metadata schema between digital libraries. After describing and identifying metadata schema, using a DTD that specifies the various attributes expressed in XML and the CNRI Handle System for schema identification, they used the DTD to develop a framework geared towards making metadata instances, schema and services into first class network objects. Using CNRI's Digital Object Architecture, these digital metadata objects were then deposited in data elements and given an Interoperable_Metadata content type regulated through a metadata registry to enable dynamic metadata conversion. As is apparent, the process is complex and there are issues of scalability. The process also requires development of software modules for each schema.

The Open Archives Initiative (OAI) develops standards and protocols for metadata harvesting to achieve interoperability between its data providers. Its use of "unqualified Dublin Core as the common metadata set was based on the belief that the common metadata set in OAI is explicitly purposed for coarse granularity resource discovery. Community-specific description, or metadata specificity, is addressed in the technical framework by support for parallel metadata sets. The technical framework places no limitations on the nature of such parallel sets, other than that the metadata records be structured as XML documents, which have a corresponding XML schema for validation" (Lagoze, 2001). While the OAI approach is useful in some contexts, its use of unqualified DCMES and retention of all optional elements means that the resultant interoperability is surface level and not suitable for the government intranet.

The authors were motivated to help the agencies involved in this case study avoid the difficulties reported to be associated with post-hoc harmonisation.

In dealing with the human and practical problems of people committing to metadata implementation, Arms (2002) adopts the term "levels of interoperability". He argues efforts to enhance and enforce interoperability can be seen as a balance between the cost of acceptance and functionality. He argues that "if the cost of adopting a standard is high, it will be adopted only by those organisations that truly value the functionality provided" (p. 4).

From these results and their own evidence of user behaviour, the authors argue that unless the process of developing interoperable metadata is simplified

and made clear to collection owners, along with the benefits, there will be problems with take-up despite the technical research. They argue that the current problems in developing interoperability solutions, identified in the literature above, have as their basis the multitude of variations to be found in most metadata records. These variations, in the case study in particular, include local adaptations of standardized metadata, local terminologies and alternative spellings and words, as well as trivial errors of use and of grammar and spelling.

## 3. Case Study Overview

Although the case study reports work that involved only a few metadata records, it did involve the future of the whole-of-government intranet and all departments of government. The problem was, how could such a wide audience be encouraged to engage with the existing problems in their metadata implementation. They were soon to be involved in extensive metadata creation for all government documents but were already questioning, after a number of years of working with AGLS metadata, the effectiveness and expense of the process. Departmental information managers were involved in the process described.

It was originally assumed that the fundamental problem was not in the process but maybe in the commitment to it. The process described showed that it was indeed the process, but that by reducing the effort and clarifying the process, government data managers could take a more active role in the production of interoperable metadata and so, in turn, achieve improved results in resource discovery and management.

The reported project aimed to achieve the following:

- Find ways to illuminate the current limitations in interoperability resulting from existing metadata practices;
- Articulate the cause of the problem;
- Develop a shared strategy for improving the interoperability, and, as it emerged,
- Encourage data managers to develop a single, comprehensive metadata application profile, derived from the current requirements and foci of all users, that does not place limits on high level local specificity but enables deep and comprehensive metadata interoperability across the particular participant group.

The result has been increased interest in harmonisation of the metadata, and the development of a shared, more detailed application profile (so far, for the six most commonly-used elements).

The on-going project aim is to help government data managers achieve complete and deep interoperability. This may be achieved now through the development of a single application profile based on existing records that incorporate metadata specific to agencies within a framework that can be accessed by all. Individual agencies might choose to operate with subsets of the application profile, in the knowledge that their application profile is fully harmonised with those of all other participating agencies. In addition, control of vocabularies and formats for metadata values has been recognised as important for interoperability, and this will be increased. The current proposal is for collaborative extension of the original AGLS profile, with greater specificity to suit the needs of the local state government.

## 4. Making Interoperability Visible - the ARH process

In making interoperability visible, the authors' approach is to *aggregate* all metadata elements from the resource collections, consider the processes that could be used to *rationalise* the aggregated set of elements and then show how the agencies might work together to *harmonise* the resulting application profile. This process is referred to as ARH – HA!: visualise the processes of aggregate, rationalise, and harmonise in order to be motivated to harmonise commonly-owned, distributed, heterogenous metadata collections.

Step one, the aggregation stage, involves the collection of data, and analysis of element usage and variation. During this stage all collected metadata tags are added to a table or spreadsheet. Any discernible variations in element names, formats or values that could confuse a search engine, such as different spellings and alternative element names and qualifiers, are recorded separately. At this point, all the differences in the use of elements are made visible and it is a simple step to seeing that interoperability could be enhanced by adding qualifiers to increase conformity and define specificity. While this may increase interoperability, it would not lessen the number of element types, or simplify the application profiles in use.

Step two is consideration of the rationalisation of the metadata. This step involves careful examination of the different metadata elements looking particularly for unnecessary variations, such as when the same value is contained in elements with different names (and namespaces) or when the same elements contain different types of values, such as different date formats. This process makes it easy to see the possibility of considerably lessening the number of types of elements, and so simplifying the application profiles and increasing interoperability.

Step three is the harmonization of the metadata. To ensure that metadata operates as a powerful and accurate communications instrument for all resources from all agencies and departments, data managers consider the use of elements and decide on harmonised approaches to their use in order to develop a shared application profile. As they agree on for-

mats or vocabularies, they see the number of elements deployed across the agencies reduced.

Functional success of the three-step visualisation process is measured by whether or not those who participate in the process do commit to harmonising their application profiles, whether it becomes ARH-HA!

## 5. ARH and Victorian government resources

Victoria is a state in the federation that is Australia. The Government of Victoria was an early adopter of the DC approach. More recently, government agencies have been attempting to improve access to their records and public documents through the use of Australian Government Locator Service application profile (AGLS) DC-based metadata. Government policy states that all Victorian government agencies should use the AGLS application profile to describe web-based online information resources. This is in line with the Federal Government's metadata directive to its departments. To date, the Victorian policy is, however, advisory rather than prescriptive. Agencies have been, to greater and lesser degrees, left to their own devices - to 'go it alone'. In fact, in the absence of any guidance other than the central policy, adoption has been spotty and often confined to what might be described as web 'brochure-ware'. Deep adoption of a unified approach to metadata has been difficult to achieve although it is now required.

In practice, departments and their agencies have used AGLS metadata and customised their application profiles, more by implication than design. Different departments use metadata for different sets of resources ranging from online, web-style public resources (classified as brochure-ware by the authors) to all resources including those embedded deeply within databases and document management systems, and never intended to be widely accessed. In addition, departments differ in their use of metadata, some seeing it as possibly useful for export to those who may need to know of the department's resources and others using it to drive their internal resource management systems.

Recently, one department has been given responsibility for developing a whole of government intranet and another for developing a whole of government public 'brochure-ware' gateway. Working on the intranet, the authors have been concerned about how to achieve high levels of interoperability of government resources. They developed the ARH activity in the process of tackling their own concerns, conscious that they were also providing a better framework in which the other agency might develop the public gateway.

First, the authors decided to test the interoperability of existing metadata records. This had been done before but it had never led anywhere. Nevertheless, a series of requests were made to each participating agency, starting with a copy of their application profile, then for sample records showing the use of the profile, and finally for a set of metadata records for analysis. This last request was made when the authors decided to experiment with the ARH process to provide a concrete demonstration of interoperability across the different sets of metadata. In all, 29 records were obtained from six of the participating departments.

The first step, aggregation of the metadata was done by creating a spreadsheet of all the records provided in order to determine variations in the metadata. To approximate the requirements of machine based searching, any variations in element names or format were treated as different elements. This was also applied to value strings where these would be interpreted differently.

Figure 1 shows a small section of the resulting spreadsheet demonstrating the kinds of variation that immediately became visible.

| Rec no. | DC.Identifier | DC.Title | title | DC.Creator | DC.Creator.nameCorporate | DC.Publisher | DC.Subject | DC.Keywords | subject/keywords | DC.Description | description | DC.Language | DC.Lang | DC.Date.created | DC.Date.modified |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | http://www.vic.gov.au | Victorian Government home | | DPC | | State of Victoria | Victoria; Victorian Governm | | Victoria; Victorian Governm | The Victorian Governm | The Victorian Governm | en | | 1999-07-20 | 2001-03-2 |
| 2 | | Department of Justice | | Department of Justice | | | Justice Consumer Affairs | | | The Department of | | | en-au | | |
| 3 | CD3EEAECE5E98E1CCA256 | Marriages (level 2 overview) | Department of Justice - | Department of Justice | | Department of Justice | 24 hour help | | 24 hour help | | | | en-au | 2002-02-01 | |
| 4 | 8BAFAF7D34AD535 | Marriage Certificates | Department of Justice - | | | | | | certificate, marriage, single | | | | | 2002-16-02 | |
| 5 | 1000151 | Fishtank | | Richard Billingham | | | | BETAVILLE, EXEMPLA | | A high-rise flat in the U.K. | | English Regional accent | | 1998-01-01 | |
| 6 | | | Department of Education | | State of Victoria, Departme | | | | education departme nt, | | Education Victoria is the | en-AU | | 2001-10-09 | 2001-12-2 |

**Figure 1. Sample of the Aggregated worksheet**

| Element name | Examples of values provided |
|---|---|
| DC.Title | Victorian Government home page |
| | Department of Justice |
| | Marriages (level 2 overview) |
| | Marriage Certificates |
| | Fishtank |
| DC.TITLE | SOFWeb Front Page |
| DC.title | Victorian Education Channel |
| title | Department of Justice - Births Deaths and Marriages - Marriages |
| | Department of Justice - Births Deaths and Marriages - Marriages - Marriage Certificates |
| | Department of Education & Training |
| | Victorian Curriculum and Assessment Authority, Australia |
| | Arts Matters |
| | Copyright, Trade Marks And Disclaimers |
| | Victorian Education Channel - Welcome Page |

Figure 2, deliberately out of focus as if seen from a great height, shows graphically the 'spottiness' of the metadata with 49 different metadata elements being used (across the top), more and less, and in a wide variety of ways (see below).

This table, displayed in full and used to illustrate the process of searching, was a useful tool. Participants posed search queries and looked at what a machine would discover. From only twenty-nine records, 49 different metatags were generated. These results made visible why a normal search across these records would produce inferior results.

In fact, variations noted within the metadata by participants included:

*Element Name Variants*

- Inconsistent case: eg. DC.Title/TITLE/title; EDNA.Userlevel/UserLevel
- Non-standard names: e.g. DC.Keywords
- Non-standard qualifiers: e.g. DC.Description.Abstract
- Non-standard abbreviations: e.g. DC.Lang

*Field Selection*

- Standardised v non-standardised element names: e.g. use of 'description' v DC.Description
- Use of created metadata names: e.g. Custodian

*Value string Variants*

- DCMES suggests certain type of value strings be used for each element/qualifier, to assist search engines
- DC.Identifier: URI recommended, other identification numbers given without qualifiers
- DC.Date: Recommended ISO8601 standard uses yyyy-mm-dd. Other formats used include yyyy, yyyy/m/d, yyyy-dd-mm
- DC.Format: Controlled vocabulary recommended a) Non-standard terms used e.g. VHS (PAL) b) Incorrect case e.g. text/HTML
- DC.Language: DC recommneds RFC1766. Variants include en, en-au, en-AU
- DC.Type: controlled vocab recommended. Non-standard Types used e.g. references and materials
- EDNA.Version: reserved for version of EdNA Metadata Scheme
- Qualifiers embedded in values: e.g. DC.Publisher CONTENT="corporateName=State..." v DC.Publisher.nameCorporate=

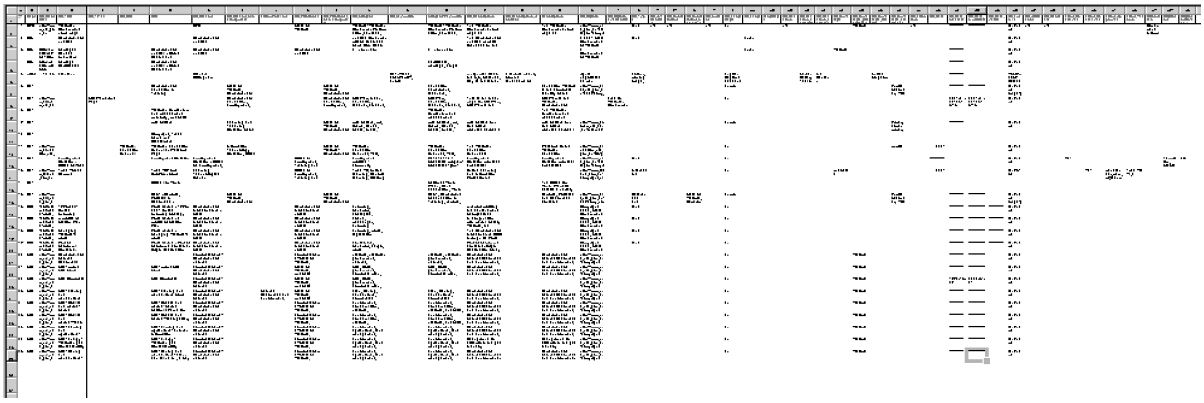


**Figure 2.**

- Non-standard proper names e.g. DPC v Department of Premier and Cabinet
- Generally inconsistent use of capitalisation and punctuation

*Overall*

- Most element variants due to non-standardised application of capitals, punctuation, spelling
- Users seem unaware of Application Profiles: Little use of collection specific qualifiers to enhance specificity

The results were then *rationalised*. The authors considered ways to reduce the size of the resulting metadata set without detracting from functionality. Elements distinguished only by grammatical errors and variations or spelling can be merged; elements not used across the different collections can be removed without loss, while those with variations of the same name (such as abbreviations of element names) can be merged. For instance, DC.Title, DC.TITLE and DC.title can all become DC.Title in HTML or dc.title in RDF. This reduces the original 49 separate tags to 42.

The final *harmonisation* process is a more complex task as it involves identifying opportunities to make decisions about best practice in the use of metatags. The first step is to look for chances to merge non-standard elements into the closest related standardised ones, e.g. merging DC.Keywords into DC.Subject. This can also apply to the use of non-standard qualifiers. Search tools, designed to access existing variations of the metadata across the various collections, can then use the merging and mapping processes identified at this stage.

Significant variations in the Victorian government metadata records were found in the format and selection of value strings and content. In describing application profiles, Heery and Patel (2000) state that application profiles may "specify permitted schemes and values" such as a particular controlled vocabulary or item string format. Thus the application profile might specify a format to be used for DC.Source instead of free text. In harmonising the records provided, each element value string needs to be addressed separately. The precise form and detail of each needs to be addressed by focus groups to incorporate the necessary input and ownership of the different stakeholders. In some cases, however, there are established practices for DCMES that can be followed more precisely, such as the use of ISO8601 for date fields.

The resulting metadata set, or harmonised application profile should then allow the specific detail of individual metadata collections to be shared and accessed by other departments and users. This could be achieved by providing clear information about the application profile in a shared registry. Given such a registry, in the future individual agencies could select from established elements and qualifications or contribute finer grained qualifiers of use without loss of interoperability. (The department representatives have now indicated their interest in establishing such a registry).

## 6. Detail of Case Study Methodology

**Size of the sample:** Seven departments were originally contacted with a request for ten records each within only three weeks. In fact, initially only five departments replied and only two of these provided 10 items. By the time that results were compiled, six departments had supplied a total of 29 records. While this was not a statistically significant sample compared with the number of metadata records owned by the departments, it proved sufficient to provide a demonstration of the process.

**Quality of the sample:** The departments were given few guidelines on what to provide for the activity. The request was simply to send examples of the documents and associated metadata from their website or intranet. While one department sent a broad range of documents representing different sections of the department, others sent information sheets or technical papers. Most records described 'brochureware' and it was noted that the associated metadata was fairly brief.

**Quality of the metadata:** As mentioned, there was a wide diversity between different documents and their associated metadata. Of the twenty nine records, 11 could be regarded as comprehensive (ie with at least 9 separate metatags), 3 were very brief (fewer than 4 tags) while 15 were between these. It was interesting to note that even in such a small sample, a wide diversity of styles was apparent.

**The rationalization process:** Selection of the criteria for rationalization caused some discussion among the participants. While those used seemed to be logical, it was agreed that there was an element of subjectivity involved and these criteria might vary based on the particular samples provided.

**Harmonization:** The harmonisation process has not yet been completed. It involves focus groups of collection owners meeting to agree on appropriate and useful metadata based on their specific needs balanced against the aims of interoperability. What is important is that the departments have agreed, after participating in the process, to work together on this harmonisation process. In one sense, it is as if the metadata process is being started afresh. This is not the case. Participants who have large collections of metadata are meeting to iron out wrinkles that have developed over time, and this activity is able to draw on five years' experience with metadata creation and use. It is better-supported by this experience than was the first attempt, and it comes at a time when a powerful outcome motivates it. Whole-of-government interoperability is no longer expected to be achieved by letting agencies work independently and hoping that technologies can be developed to reintegrate the

metadata post-hoc. The need for planned interoperability has become visible and is now being made operational.

## 7. Conclusion

At a meeting of information managers from the government departments working together on the intranet, collection owners expressed satisfaction with the results of the work done so far. One stated that the display (visualization) of the interoperability of the current metadata had made her see the importance of standardization. Another said that from now on his agency would increase its efforts to generate more useful metadata for their collection.

The final application profile has yet to be delivered but it is anticipated that it will be accepted much more readily than previous profiles because of the local input in developing it. The authors conclude that this process would not have been undertaken in the context if the ARH process had not been developed and attracted the managers' participation. Particularly as it was not the first attempt to achieve the outcome, but was successful. Further, the authors recommend the activity as being useful to those working with information managers and others who are developing practices and implementing established application profiles. The visualisation of interoperability seems to be useful in such a context.

## References

Arms, W et al. 2002. 'A Spectrum of Interoperability: The Site for Science Prototype for the NSDL'. D-Lib Magazine 8(1) Jan 2002. [Online] http://www.dlib.org/dlib/january02/arms/01arms.html [Accessed 2002-06-12].

Bearman, D. et al. 1999. 'A Common Model to Support Interoperable Metadata'. D-Lib Magazine, Jan 1999 [Online] http://www.dlib.org/dlib/january99/bearman/01bearman.html [Accessed 2002-06-13].

Blanchi, C. & Petrone, J. 2001. 'Distributed Interoperable Metadata Registry'. D-Lib Magazine 7(12) Dec 2001. [Online] http://sunsite.anu.edu.au/mirrors/dlib/dlib/december01/blanchi/12blanchi.html [Accessed 2002-06-10].

Heery, R., Patel, M. 2000. 'Application Profiles: Mixing And Matching Metadata Schemas', Ariadne Issue 25, Sept 2000 [Online] http://www.ariadne.ac.uk/issue25/app-profiles/ [Accessed 2002-06-13].

Hunter, J. 2001. 'MetaNet - A Metadata Term Thesaurus to Enable Semantic Interoperability between Metadata Domains'. JoDI 1(8) February 2001 [Online] http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter/ [Accessed 2002-06-16].

Lagoze, C. 1996. 'The Warwick Framework: A Container Architecture for Diverse Sets of Metadata', D-Lib Magazine, July/August. [Online] http://www.dlib.org/dlib/july96/lagoze/07lagoze.html [Accessed 2002-06-16].

Lagoze, C.& Van de Sompel, H. 2001. 'The Open Archives Initiative: Building a low-barrier interoperability framework'. [Online] http://www.openarchives.org/documents/oai.pdf [Accessed 2002-06-19]

Weibel, S., Koch, T. 2000. 'The Dublin Core Metadata Initiative'. D-Lib Magazine 6(12) Dec 2000. [Online] http://www.dlib.org/dlib/december00/weibel/12weibel.html [Accessed 2002-06-16].

## URIs

Open Archives Initiative. http://www.openarchives.org/

The Dublin Core Metadata Initiative. http://dublincore.org/

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 185-192*


# Metadata Pilot at the Department for Education and Skills, UK

Julie Robinson
Assistant Librarian, Library and Information Services Team,
Department for Education and Skills, London, United Kingdom.
Julie.robinson@dfes.gsi.gov.uk

## Abstract

*This paper describes the Department for Education and Skills' (DfES) practical approach to tackling metadata and surrounding issues. A metadata pilot project was set up by the Library and Information Services Team to develop a metadata scheme for departmental use. Using the Dublin Core based e-Government Metadata Standard (e-GMS), Library staff developed a draft metadata standard for departmental web pages. Library staff applied the metadata standard by metatagging pages on a test web site. The metatagged pages were tested against the search engine. Work started on the pilot in September 2001. The pilot was successfully completed in November 2001. Further developments are ongoing.*
**Keywords:** *Dublin Core, Metadata, Department for Education and Skills, DfES, interoperability, e-Government Metadata Standard, e-GMS, Government metadata.*

## 1. Introduction

The Department for Education and Skills is an important, central UK Government Department (www.dfes.gov.uk). Its mission is to create opportunity, release potential and achieve excellence in education for all. We rely on the departmental Intranet as an essential communication tool. Metadata is an important part of this tool that will improve the ability of all staff to retrieve the information they need. This is necessary to improve the delivery of public services which is a key goal of the broad agenda to modernise UK Government.

The DfES Intranet has over 100,000 web pages for 4,000 staff, nearly all Civil Servants, who work in four separate locations across the UK. When staff search on the Intranet using the internal search engine, they often have difficulty finding what they are looking for. In addition, searchers are often presented with too many results pages, many of which are only slightly relevant to what they are looking for. Using the Dublin Core based draft e-Government Metadata Standard (e-GMS) as a starting point, the metadata pilot was set up with the purpose of investigating how this problem might be remedied.

## 2. Search and retrieval issues at the DfES

In the DfES, web pages are produced for the Intranet and for the Internet site. The Intranet consists of over three hundred web sites. These are all sub sections of the intranet. Effective responsibility for these sites is devolved to 200 web managers. There is little editorial control and web managers are expected to publish according to prescribed web standards. Prior to the pilot, there was no policy of using metatags and only a handful of web managers used them. Internet web pages are published by six well-trained web developers who had just started to add metadata to web pages when the pilot was set up.

Although the Intranet search engine was configured to search in a standard way (i.e. using titles and keywords as indexing terms automatically produced by software agents), the search results produced were often not very relevant for users. The rankings also tended to be questionable. During the pilot, it was found that many web pages did not have meaningful titles and that many still had the default title (i.e. no title). Library staff had worked with Intranet and IT colleagues to successfully redesign the Intranet search interface to help users but these problems still remained. One of the reasons why library staff set up the metadata pilot was to investigate how metadata might solve these problems.

## 3. Establishing the metadata pilot

The main drivers behind the pilot were:

- the mandatory requirement to make all government services available electronically by 2005 (81)

- the publication of the UK e-Government Metadata Framework (e-GMF) in May 2001.The e-GMF set out the UK government's policy for standardising metadata use throughout the public sector. (92)
- the circulation of the draft e-GMS in month September 2001. (63)
- the establishment of the DfES extranet project in the Summer of 2001.

Initially, the recently formed extranet development team asked library staff to supply a single metatag to control data transfer from the intranet. (An extranet project had just been set up and the DfES extranet was subsequently launched in April 2002. It links the Department up with its external governmental partners). However, library staff were also aware of the need to add metadata to public sector resources as mandated by the e-GMF which adopted Dublin Core as the UK Government Metadata Standard. Dublin Core was adopted because it is a highly developed, flexible, internationally recognised model. The e-GMF set out the UK Government's policy for standardising metadata use throughout the public sector, and has since been superceded by the e-government Interoperability Framework (e-GIF) v4.(10) This is complemented by the e-GMS which describes the elements and their refinements. Once a draft version of e-GMS was available in September 2001, a sound basis existed for establishing a metadata pilot to test a range of metatags(11)

## 4. Running the pilot

In September 2001, as a pilot project, library staff set created a metadata schema and a draft metadata framework for departmental use according to the e-GMS. Library staff applied the draft DfES framework, using the metadata scheme, to a test site on the intranet. All pages were metatagged appropriately by the end of October 2001. The metatagged pages were then tested against the intranet search engine. The pilot was successfully completed by early November 2001.

The e-GMS was devised because Dublin Core alone is not sufficient to meet all of the government's information management and information retrieval needs e.g. records management and data security. To meet these purposes, the e-GMS therefore added further elements and refinements whilst following the principles of Dublin Core. That said, the e-GMS is not a one size fits all standard. Local metadata standards, consisting of sub-sets of the e-GMS, need to be developed to meet the specific needs of any given organisation. Thus the need to create a draft DfES Metadata Standard as part of the pilot.

**Table one. Metatags, HTML view.**

A crucial aspect of exploiting added metadata is that the search engine needs to be configured to enable field searching. For this, specialist advice will have to be sought.

Library staff decided to add metadata directly to the web pages in HTML. This was the quickest and easiest way of adding metadata for the purpose of the pilot. The other main advantage of this method is that is inexpensive (i.e. we did not have to purchase metataging or content management software).

```
<html>

<HEAD>
<!— MetaTager : 0001S —>
<meta name="AUTHOR" content="none">
<meta name="TITLE" content="none">
<meta name="DESCRIPTION" content="none">
<meta name="SUBJECT" content="none">
<meta name="IDENTIFIER" content="http://
ntweb1/">
<meta name="DATE.CREATED" content="none">
<meta name="DATE.LAST_UPDATED" content=
"none">
<meta name="DISPOSAL.REVIEW" content=
"none">
<meta name="ACTION Archive" content="none">
<meta name="RIGHTS.BUSINESS_GROUP
_ACCESS_PERMISSION" content="PUBLICDO-
MAIN">
<meta name="ALTERNATIVE TITLE" content=
"none">
<meta name="AUDIENCE" content="none">
<meta name="CATEGORY" content="none">
<meta name="CONTRIBUTOR" content="none">
<meta name="COVERAGE.PLACE" content="none">
<meta name="FORMAT" content="Web site">
<meta name="KEYWORDS" content="none">
<meta name="LANGUAGE" content="Eng">
<meta name="PRESERVATION" content="none">
<meta name="PUBLISHER" content="none">
<meta name="RELATION.ISBASEDON" content=
"none">
<meta name="RELATION.ISPARTOF" content=
"none">
<meta name="RELATION.ISVERSIONOF" content=
"none">
<meta name="RIGHTS.COPYRIGHT" content=
"Department for Education & skills. www.dfes.gov.uk/
disclaimer.shtml">
<meta name="TYPE" content="Text">
<meta name="TYPE.DOCUMENT" content="Web
Page">
<HEAD>
```

The pilot established four main entry points for searches; author, title, subject and keyword, and established that a special query language had to be used to search on metadata. Finding a method that allowed individual tags to be searched was the difficult part of the pilot. Microsoft, the software provider, produced a guidance listing query language.

However, this was not accurate or complete, being a standard guide and not one for the DfES intranet, so the query language for the tags had to be worked out by trial and error, by testing the method on a few pages set up for this purpose (initially half a dozen). This involved adding and removing meta content and tags, changing the syntax and then running controlled searches. Using unusual search terms helped library staff to do this to confirm that the tags were working. (These terms were later removed).

The result of this was that, initially, the metadata and the syntax used had to be configured to work with the search engine. It should have been the other way round. (IT colleagues later successfully configured the search engine to work on metatags specified by library staff in the metadata scheme).

## 5. The metadata schema (12)

The short version of the DfES Metadata Schema looks like this:

**Table two. DfES Metadata Scheme. Pilot version**

The tags were chosen and described according to the e-GMS and the DfES Metadata Framework by Library staff based on our knowledge of the information needs of the Department. We were also aware that, if the metadata scheme were to be widely adopted, it would need to be as simple and easy to apply as possible. These tags were based on the e-GMS current at the time. The standard has since changed and the DfES metadata scheme has changed accordingly. Lack of space prevents a complete discussion of all the elements used, so discussion is based on key issues.

### 5.1 Author

This should have been "Creator" to conform to Dublin Core and the e-GMS. However, the search engine did not work on 'creator', only 'author'. It would seem that the default metatags recognised by the search engine software included "Author" not "Creator". The search engine was later configured to recognise and use "Creator" which is part of the current metadata scheme.

### 5.2 Description

The description provided should help users identify the right information in a list of search results. It should also help users identify web pages they are not looking for the information in a list of search results. Library staff wrote the descriptions based on a reading of the resources in question and a familiarity with the test website and the likely needs of users. Seeing that the resource is not relevant immediately saves users/searchers time and prevents them from getting the wrong information or information that is not required. The searcher reading the description (or abstract as it is called on the Intranet) should be able to tell if the page is worth reading from the description provided without having to go into the page itself and wade through the text. (13 )

### 5.3 Subject

The e-GMS allows Keyword and Category as refinements of Subject. The Category refinement is to be used for terms from the Government Category List (GCL) to aid cross- government browsing.

### 5.4 Subject (unqualified)

This is a very useful metatag tag because the search engine can pick this up and match it to the search terms entered by someone searching on the Intranet. Library staff used terms suggested by the test site owners supplemented by some of their own choosing. These included buzz words and phrases like "box times". (This is the daily time a document needs to reach a Minister's office to ensure the Minister sees it that evening. Box times are central to our working practices, and they change, particularly during the parliamentary recess). Staff also included abbreviations, e.g. PQs as well as the full term "parliamentary questions". This facilitates better resource discovery. Terms were sometimes suggested by resource content e.g. "ministers' responsibilities". This allows users to find very specific information quickly. These terms are all uncontrolled. This means that they were freely chosen and not limited to a prescribed set. There are no restrictions on the definition or usage of such terms.

### 5.5 Keywords

These are subject terms but put in a different tag because these terms are all drawn from a controlled vocabulary, the Departmental thesaurus. So we have "Prime Ministers" as a keyword, but "Tony Blair" is a subject term. By combining search terms in this way, we have introduced some synonym control. The important thing is that whatever the search term entered, the resource produced in the hit list should meet the user's needs. Linking search terms in the free text element and the controlled vocabulary will facilitate this discovery. Searching on keywords helps to reduce excessive numbers of hits. This is an important advantage of having a controlled vocabulary.

### 5.6 Action (n.b. not part of Dublin Core)

This is for archiving purposes and library staff set this value to "Archive" because the test site was considered to be of intrinsic historical value and interest. We would expect that it would be kept, possibly even-

**Table 1.**

| Element Name | Refinement | Definition |
|---|---|---|
| AUTHOR | | Person, group or organisation responsible for the intellectual content of the resource. |
| TITLE | | The name given to a resource. |
| DESCRIPTION | | A description of the information contained in the resource. |
| SUBJECT | | Uncontrolled key words and phrases indicating the subject matter of the resource. |
| IDENTIFIER | | The unique identifier of the resource (the URL or web page address). |
| | DATE.CREATED | The date the resource was created. |
| | DATE.LAST_UPDATED | The last time a resource was updated or altered. |
| | DISPOSAL.REVIEW | Date on which the resource should be reviewed to determine the need to retain it. |
| | DISPOSAL. ACTION | If the resources has a long term value. |
| | RIGHTS.BUSINESS_GROUP -ACCESS_PERMISSION | Defined groups to which access to the resource is limited. |
| ALTERNATIVE TITLE | | Any alternative name or title by which the resource may be known. |
| AUDIENCE | | The target audience of the resource. |
| | SUBJECT. CATEGORY | Key words and phrases indicating the subject matter of the resource taken from the Government Category List. |
| CONTRIBUTOR | | The person or organisation that has played a part in creating the resource but does not appear in the author element. |
| | COVERAGE.PLACE | This is place covered by the content of the resource. |
| FORMAT | | This is the physical format of a resource. |
| | SUBJECT.KEYWORDS | Key words and phrases indicating the subject matter of the resource taken from the Departmental thesaurus. |
| LANGUAGE | | The language of the data of the resource. |
| PRESERVATION | | Data needed to support the perpetual preservation of the resource. |
| PUBLISHER | | The organisation a user needs to contact to obtain permission to re-publish the information contained in a resource or to obtain copies in a different format. |
| | RELATION.ISBASEDON | The resource is an adaptation, translation, derivation or interpretation of another resource. |
| | RELATION.ISPARTOF | This is when the resource is a physical or logical part of another. |
| | RELATION.ISVERSIONOF | The resource is a version, edition or adaptation of the referenced resource. |
| | RIGHTS.COPYRIGHT | Indicates the User's rights to view, copy, redistribute, republish or otherwise make use of all parts of the resource. |
| TYPE | | This relates to the genre or category of the resource. |
| | TYPE.DOCUMENT | This relates to kind of information contained within the publication. |

tually going to the Public Record Office (PRO). The Public Record Office is the National Archive of England, Wales and the United Kingdom. It brings together and preserves the records of central government and the law courts, and makes them available to all who wish to consult them.

**5.7 Date review**

Library staff varied this using common sense as to when a reasonable person could or would expect the resource to be updated. UK constitutional requirements for the holding of elections, mean that general elections must be held within five years maximum from the date of the last election, and government web sites need to remove items from the previous administration and replace them with web pages describing the new administration. Even if the same political party wins the next election, the personnel of Government will undoubtedly change and this will need to be reflected in the Department's web content. The review date is therefore often set this to 4 years from date of last election. Again, further guidance would seem to be in order. However, this element does presuppose a content management policy.

**5.8 Relation.Is Part of**

This gives the URL for the test web site Home Page. This will facilitate resource discovery because it helps the search engine to identify and pick up relevant pages in many cases. Most search engines only skim the surface of a site. However, the really useful, more information rich resources are often located on a deeper level (four or five clicks away). This means they are often not picked up leading to poor results. By filling in this element, retrieval of relevant resources is improved. This is because the metadata links materials/pages thereby producing better search results.

**5.9 Rights: Business Group Access Permission**

This metatag was included to meet the requirements of the extranet project. This means that it determines if a webpage can be copied over from the intranet to the extranet. Given the presumption of openness which should apply in an open system of government, library staff suggested the default should be "Public Domain". The meaning was of this term was clarified in the Department's metadata framework and guidance produced. Later this was simplified to "Public".

## 6. The benefits of using metadata

The benefits of metadata fall into two categories: searching and other benefits.

**6.1 Searching (14 )**

The main points to note are:

- Doing searches using metadata produces better search results. Much peripheral or irrelevant material eliminated and the results were noticeably more relevant.
- The quality of the abstracts is an improvement on machine generated descriptions which often do not make sense. This saves user's time by facilitating the quick evaluation of results.
- Metatagged items are ranked higher by search engines, so retrieval of relevant items is improved.
- There are fewer hits in the results lists. Non-relevant material is greatly reduced and precision is improved. There are no false drops.
- In an age of information overload, less is more. If metadata is not used, time creating valuable resources is wasted because they cannot be found or are lost in an overload of "hits".

**6.2 Other benefits**

Any system produced for one reason will tend to have knock on benefits for other, sometimes unintended, purposes. Metadata is no exception.

The main points to note are:

- It has highlighted the importance of web standards. For example, when doing test searches, library staff noticed that web bots sometimes came up. (Web bots are components of a Front Page Web page that simplify development tasks e.g. an organisational logo). This was because they had not been placed in a private folder where they could not be searched.
- It adds value to resources by adding information not always available in the resource itself e.g. author and date of publication. We take this information for granted with paper resources. However, the ease of web publishing has come with the disadvantage that it often lacks metadata. This is important because metadata adds to our knowledge of the provenance, currency and reliability of web based information resources.
- Content management is enhanced through the review and date tags. This information can be used to keep sites and information accurate and up to date e.g. it is possible to auto generate email to authors to update documents. This has the added advantage of making authors take responsibility for their documents once published. The DfES is currently working on this.
- The preservation tag (n.b. not part of Dublin Core) can be used for records management purposes and may be useful for electronic document and records management systems.
- Useful resources of long-term value can be identified. This avoids duplication of effort and the loss (and costly replacement) of information rich resources.

- It can be used as a tool to facilitate data/resource transfer. The access tags indicate which resources can or should be transferred (and which not).
- Finally, it increases awareness of the importance of information as an asset and its value.

## 7. Disadvantages

Despite the advantages gained from adding metadata, there is a price to pay in terms of some disadvantages. The main ones are:

- Metatagging does take time, irrespective of who does it.
- Metatagging on a wide scale will cost money. If specialist software is used, this could add to the costs.
- Implicit in our approach is the assumption that tagging will be widely devolved to authors rather than done by a small number of indexing professionals. This means that it may well be difficult to maintain the necessary standards required to gain the full benefits of applying metadata.

## 8. Outcomes

The main outcomes of the pilot were:

- A rights tag was created to successfully meet the requirements of the extranet project.
- There is a DfES metadata framework for web pages based on the e-GMS, which conforms to Dublin core.
- There is a short guide for web managers on how to metatag web pages.
- The intranet search engine now has much greater functionality having been configured to recognise and use metadata and use it in searching.
- It was decided to implement Metadata across the Department as part of the extranet project. The extranet was launched in April 2001 and metadata is gradually being added to key sites as part of a rolling programme.
  Some changes have been made since the pilot e.g. the metadata will be input via a web authoring tool not directly via the HTML and the metadata scheme was amended and improved by the addition of suitable encoding schemes.

## 9. Next steps/challenges

Staff need to re-write the intranet search interface to allow metadata to be used by the search engine without users needing to know a special query language. Users also need the option to search explicitly on metadata as an advanced option, again without using a special query language or knowledge. This will require further resourcing and development work than was initially anticipated.

The Department will launch a portal later in 2002. The new search engine (Verity), the portal software (Plumtree) and the categorisation software (Semio) will need to be harmonised and configured to use metadata. This will be a large and complex task. The main problem here is that metadata cuts across many aspects of the portal and therefore presents a hurdle in terms of co-ordination.

Library staff will need to revise and update the DfES Metadata Standard in the light of the above. Here, the main problem is coping with a moving target.

More metadata frameworks/standards are needed, especially for word processed documents. This raises the question of whether to have one overarching standard for all formats or to have one standard for each format. The former could be unwieldy whilst the latter approach might lead to confusion on the part of users and authors.

The e-GMS will be updated later in 2002 to take into account the PRO's Records Management Metadata Standard and other requirements that have come to light. This may lead to the DFES Standard being edited. This raises the problem of making changes to standards in a controlled manner and then ensuring that the new standard is understood and implemented. This raises the issue of how to comply with both the e-GMS and the PRO's forthcoming Record Management Metadata Standard.

Methods are needed to make it easier to add metadata to resources, especially adding keyword terms from the Department's thesaurus. The main questions here are how to get staff to metatag and how to deal with the resource in terms of time, staffing and the cost that this involves. Getting users to add keywords from the thesaurus is a particular concern as our experience as information professionals shows that few users use this facility. Yet this is vital to gain the benefits of a controlled vocabulary.

We will need to think carefully about how to expand this pilot to the forthcoming Electronic Document Records Management System. At time of writing, there is no recognised pan-government standard which can be used and there are a limited number of systems which can apply metadata in a way which meets the Department's requirements.

Implementing metadata implies change management. This means that the DfES corporate knowledge and information needs to be more explicitly structured, cohesive and readily accessible and that individuals must assume a greater level of responsibility for the information resources they produce. This will be difficult because it will require a change in accepted practices.

## 10. Lessons learned and conclusions

The pilot showed that it was possible to implement metadata within the DfES environment. Establishing

good working relationships with IT colleagues outside the library team was very productive. During the pilot, the Library staff gained a greater understanding of the technical aspects of metadata and established effective working relationships with IT. Library staff continue to take metadata work forward from the pilot to implementation across the Department.

We also learned that there are different ways of adding metadata than directly by html. Using the properties option in the web authoring tool and using a template are the other two methods which we realised were also possible. On reflection, library staff concluded that using a template might well be a better way to add metadata. Templates are easy for users to fill in and can be built into the workflow process. However, we do not yet have a way of using a template for this purpose.

As already noted, adding metadata can be costly and time consuming. This is an important issue which we have yet to fully address. This is important as support and compliance are vital. Added to this is the problem of how much to metatag.

We learned that itis important to ensure that the search engine is configured to be compatible with the metadata profile. Otherwise the metatags will not work.

We also learned the importance of standards. If metadata is not consistently applied, the benefits can be lost. This also shows the importance of information policies. Here the main problem is getting high level support for such an approach.

Library staff also realised the limitations of html. The fact that a single character space out of place can make a metatag fail to work properly shows that html is not sufficiently syntactically strict. Using XML (which is syntactically strict) might have produced better results.

The pilot highlighted the need for new search interfaces which will use metadata without the need for special query language or knowledge by the searcher. We are currently working with IT colleagues on developing such an interface.

For library staff an important lesson was that there is a need for a group to co-ordinate and promote metadata in the Department. At time of writing, this is under consideration.

The project also raised the issue of how diverse individual applications of the e-GMS will become. Even with the e-GMS acting as a 'master list' and giving detailed guidance on implementation, variations may begin to appear between different applications. This question cannot be properly answered until further projects are undertaken. It will be interesting to see how other UK Government departments do this and the schemas they produce.

This leads to the question of metadata registries. Although research in this area is still in its infancy, it has been noted that metadata registries are thought to be an answer to the problem of sharing and standardising internet based information. (15 )

As yet, there is no e-GMS metadata registry. However, following this concept, the UK Gov Talk website was established enable the Public Sector, Industry and other interested participants to work together to develop and agree policies and standards for e-government. This is achieved through the UK GovTalk consultation processes (http://www.govtalk.gov.uk). GovTalk will also hold a repository of XML schemas, which anyone can use to ease the implementation of new systems, this should help standardise application of metadata across the UK public sector.

Finally, it is expected that the DfES Metadata Standard will be published in due course on GovTalk as part of this ongoing process.

## Acknowledgement

The author would like to thank the following: Paula Collins, Maewyn Cumming, Anne Horan, Peter Newell, Mathew Pearce, Tracey Reeves, Sue Rolfe, Patrick Ryan, Christine Tate.

## References

1. Milstead, J. and Feldman, S. 1999. Metadata: cataloguing by any other name. *Online*, 23(1).

2. Office of the E-Envoy. 27 March 2001. UK Highlighted as global leader in electronic government. Press Release. CAB 091/01. http://www.e-envoy.gov.uk/news/pressreleases/html2001/27mar01.htm. See also: Special symposium edition (2001). "Metadata: a networked information strategy to improve access to and management of government information". *Government Information Quarterly* 18.

3. Dublin Core Metadata Initiative. Projects. Available at: http://dublincore.org/projects/. See also: http://dublincore.org/news/adoption/ for details on the Adoption of Dublin Core by governments.

4. US Government Information Locator Service. http://www.access.gpo.gov/su_docs/gils/index.html . See also: Moen, W.E. and McClure, C.R. Washington, USA. June 30th 1997. An Evaluation of the Federal Government's Implementation of the Government Information Locator Service. http://www.access.gpo.gov/ su_docs/gils/gils-eval/index.html
http://dublincore.org/projects/

5. National Archives of Australia and Office for Government Online. Version 1.2. August 2000. The Australian Government Locator Service (AGLS) Manual for Users. http://www.naa.gov.au/recordkeeping/gov_online/agls/user_manual/intro.html

6. Cabinet Office, Office of the e-Envoy, September 2001, *The UK e-Government Metadata Standard v2*, Cabinet Office, Office of the e-Envoy, Unpublished.

7. Latest Version: Cabinet Office, Office of the e-Envoy, May 2002, *GCL (Government Category List) version 1.1*, Cabinet Office, Office of the e-Envoy. Available at: http://www.govtalk.gov.uk.

8. The Cabinet Office, 1999, *Modernising Government*, Cm 4310, Norwich, The Stationery Office. http://www.cabinet-office.gov.uk/moderngov/whtpaper/index.htm

9. Cabinet Office, Office of the e-Envoy, 2001, *e-Government Metadata Framework*, London, Cabinet Office, Office of the e-Envoy. http://www.Govtalk.gov.uk/egif/home.html

10. Cabinet Office, Office of the e-Envoy, 2001, *e-Government Interoperability Framework*, London, Cabinet Office, Office of the e-Envoy. e-GIF 4 http://www.govtalk.gov.uk/interoperability/egif_document.asp?docnum=534

11. Cumming, Maewyn, 2001. Metadata in the UK. *In: DC-2001 Proceedings of the International Conference on Dublin Core and Metadata Applications 2001.* Tokyo: National Institute of Informatics, October 2001, 263-269. Available at: www.nii.jp/dc2001/proceedings/product.

12. For a discussion of metadata schemas and application profiles as a type of metadata schema see: Heery, R. and Patel, M. September 2000. Application Profiles: mixing and matching metadata schemas. *Ariadne* Issue 25. Available at: http://www.ariadne.ac.uk/issue25/app-profiles/intro.html.

13. Craven, T.C. 2001. Description Meta tags in locally linked web pages. *Aslib Proceedings*. Vol 53, No. 6, pp. 203-216.

14. One example of research on the impact of metadata on information retrieval is: Agnosti, M, Crivellari, F. and Mclucci, M. 1999. The effectiveness of metadata and other content descriptive data in web information retrieval. Paper presented to the *Proceedings of the Third IEEE Metadata Conference*, National Institute of Health, Maryland, USA, 6-7 April, 1999. http://computer.org/proceedings/meta/1999/paper-u=index.html.

15. Polyoratou, P. and Nicholas, D. 2001. Familiarity with and use of metadata formats and metadata registries amongst those working in diverse professional communities within the information sector. *Aslib Proceedings*. Vol 53, No. 8, pp. 309-324.

Posters

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 195-196*


# Policy Control Network Architecture using Metadata

Ray S. Atarashi
Communications Research Laboratory
ray@crl.go.jp

Shigeru Miyake
Hitachi, Ltd.
yake@sdl.hitachi.co.jp

Fred Baker
Cisco Systems
fred@cisco.com

## Abstract

*Quality of service (QoS) technology has been implemented to be applied to new applications on the next-generation Internet. However, as new applications such as P2P and stream application have many kinds of features and requirements, some additional features should be added to current QoS control technology. Policy definition for transport layer in a domain and among domains is being discussed at IETF to set a standard process, however detailed policy corresponding to the application or contents information according to the application semantics has not been discussed. Therefore we developed QoS policy control mechanism using metadata which is defined as a structured data according to the application semantics. Though metadata and transport mechanism can be located into quite different positions in the concept of network layers, we made them successfully collaborated by defining meta policy. In this paper, we describe our system architecture to define a meta policy based on the requirements and information contents from the application as a high level layer concept to be able to classify the network behavior. Our approach enables to multiple QoS control and collaboration among domains.*

**Keywords:** *Metadata, Quality of Service, Diffserv.*

## 1. Introduction

On the next generation Internet technology, a lot of new applications have been developed such as voice, video stream, and database transaction. New multimedia applications use various types of media with different features. Each type requires a particular quality of communication service to be transported on the Internet. For example, voice data packets should be sent in order, without jitter. The database transaction packets can be sent in order, packet loss is a serious problem, however, a short delay is acceptable.

To achieve such different transfer requirements for each data type, following technologies are introduced in last a few years. The QoS for service differentiation is called "differentiated service (Diffserv)" [1]. Diffserv architecture consists of definition of transport service classes, detection of data flows, and control of data transmission according to the defined classes. Each class of service mapped to The packets that flow on the network are classified by Diffserv code point (DSCP) [2] has different ways of regulating network behavior, defining of transmission parameters and dropping packets. The classification is put in the packet header. The regulation of network behavior is called a "policy".

On the transport layer communications, applications are identified by transport protocol, port number and a pair of source and destination addresses. QoS control is required by the applications and service types. However, applications can not control detailed QoS because decision point of QoS control is low level transport. For the deployment the QoS technology, high level QoS requirement and policy from applications should convert to low level QoS control on the transport. However, it is not considered in QoS technology.

Additionally, some application needs multiple QoS control on the connection that has same transport protocol, port number and a pair of source and destination addresses. For example, web service provide many kinds of media data on the same protocol such as text, voice, and so on. Even if the media type which program used are same, QoS requirements are different according to various those meanings. For instance, in an emergency, a lot of people communicates each other, e.g calling ambulance, communication with family and friends, and so on. These communications have priority classes and expected to transmit differently.

Most application establishes peer to peer connection across policy domains that managed consistent policy through network nodes. It is needed to make consensus about policy among domains. Since Diffserv architecture defines only packet marking and per-hop forwarding behavior on network node, it is difficult to exchange policies requested each application.

To overcome these problems, we developed a policy control mechanism by application. Our mechanism adopted metadata to describe application poli-

cies. Though metadata can be used for contents and applications to manage information, we confirmed that metadata can describe QoS policies for transport communications.

## 2. Policy Control Mechanism using Metadata

Metadata is the first level that the application provides the policy. At present, this policy described by the metadata does not control QoS functions. Our mechanism conveys application policy defined by metadata to the Diffserv DSCP. We adopt Dublin Core Metadata [5]in this paper because Dublin Core is the most popular metadata in the digital resource and its registry system is strongly needed for keep consensus among some domains.

The data and application need to be identified to the network, in order to gain service from the network appropriate to it. However, this information is usually available to the application only in its terms - object format, content, owner, and so on. What the network concerned is the type of service that the application requests. Such type of services should be expressed in terms of the local network management policy. Therefore, we need a translator.

We propose that the translator should have some form of API. It may resolve requests using a local configuration file which describe the meta policy. Meta policy select the appropriate DSCP value for the IP packet. When meta policy select DSCP value, it considers the operating system and application's attributes which are stored in LDAP directory. The application then invokes an interface method to apply the DSCP to the outbound data stream. We defined "meta policy" to translate from metadata to DSCP.

Metadata is described by resource administrator who controls application level policy. Meta policy is created by service administrator and network administrator. Service administrator gives application policy, network administrator specifies how to converts application policy to transport policy. Note that meta policy stores both application policy and transport policy at the local host computer and the local host computer specifies the DSCP value. In conventional QoS models, the transport policy is stored in the policy servers which mark the DSCP value. Since the policy servers cannot access the application policy at the local host computer, the conventional QoS models could not consider the application policy.

Fig. 1 shows system architecture which use metadata registory for keep same policy among some domains. application A sets a DSCP, and application B sets it according to a B's local meta policy.

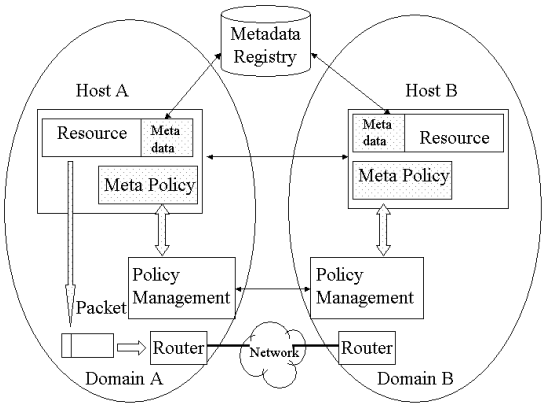



**Figure 1. Policy mechanism architecture for two domains**

## 3. Conclusion

Detailed QoS control is strongly required in the next generation Internet applications. Metadata is going to important for not only structuring and discovery digital resource but also communication interaction. This mechanism is discussing at IETF to make standard and deploy Diffserv QoS. An Request for comments (RFC) is going to publish as Best Current Practice (BCP) and start to discuss to make consensus [6].

In addition, This mechanism is going to be included in International Telecommunication Union, Telecom Standardization (ITU-T) F.706 recommendation: "Service Description for an International Emergency Multimedia Service".

## References

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, "An Architecture for Differentiated Services", RFC2475, Dec., 1998.

[2] K. Nichols, S. Blake, F. Baker, D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC2474, Dec., 1998.

[3] Takeshi Aimoto, Shigeru Miyake, "Overview of DiffServ Technology: Its Mechanism and Implementation", IEICE transactions on information and systems, Vol.83, No.5, pp957–964, 2000.

[4] Weibel, S., Kunze, J., Lagoze, C., M.Wolf, "Dublin Core Metadata for Resource Discovery", RFC2413, Sep., 1998.

[5] Dublin Core Metadata Initiative, http://www.dublincore.org/

[6] R. Atarashi, F. Baker, "Reflexive DSCP Policy", Internet-Draft, IETF, June., 2002.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 197-198*


# An Online Knowledge Gateway for Industrial Design Education and Research Activities

Mida Boghetich, Paolo Ciuccarelli, Perla Innocenti, Federico Vidari
Dept. INDACO, Politecnico di Milano, Italy
paolo.ciuccarelli@polimi.it

## 1. Introduction & issues

This poster presents the development of DesignNet, a knowledge-based project to the online digital display, retrieval and archiving of rich media resources for industrial design education and research. The project addresses the needs of end-users (teachers and students) and content providers interacting with the School of Design of the Politecnico di Milano. It moves from the assumption that conventional modalities of archiving and presentation currently adopted by the Politecnico and other academic institutions are not coherent with design teaching. Typical outputs of industrial design process are in fact 3D models or 2D graphics (digital and/or physical), not just texts or simple images. The challenges, philosophy and methodology in creating this evolving Web-based, cataloguing, multimedia knowledge-base of Virtual Reality and textual design resources are discussed, along with description of the related system and prototype.

Traditional cataloguing standards or automated search engines are not efficient with files such 3D models and 2D graphics. Their performances degrade with multimedia data, as they have not been created to be catalogued and their textual information are implicitly contained but not explicitly declared. Design knowledge is mainly iconically manifested and manipulable, thus most of present resources resulted "invisible" and did not have a defined and organized location through time. Easy usability, transfer and visualization of such data were identified as main goals of the project. Furthermore, as such resources are in large quantities and in constant evolution, we needed an open, integrated and collaborative structure, with multiple levels of description and the possibility of regular checking and updates.

## 2. The system

DesignNet gateway features a searchable, browsable database of high quality resource collections and services, recognized cataloguing and indexing standards and specific interface modalities. The application profile schema used in the project is based on Dublin Core Qualified, chosen because of its flexibility and simplicity.

Selection and definition of Dublin Core schema for the project has been experimentally led testing different schemas with content providers (departmental and interdepartmental laboratories and archives and the Permanent Design Collection of the Triennale Foundation of Milan).

The recently released OAI Version 2.0 of Protocol for Metadata Harvesting is being investigated for interoperability issues, although we are aware that this will result in a loss of the detailed qualification that has been done within the project.

In order to provide a unified integrated access, a in-house Italian Thesaurus for Industrial Design has been manually developed using terms and classes from specific domains. This term-based approach was preferred as it improves the precision for descriptions and subject access to resources, enabling more updating, exhaustiveness, specificity and flexibility than classification and subject headings. We referred to ANSI/ISO standard, Dewey Decimal Classification, manuals and pre-existing thesauri. The School of Design community was also actively involved for term selection and class organization, according to the main issues occurring in the creation of a design project.

Parallel to the thesaurus is the elaboration of authority files of companies, institutions and relevant people of the design world, periodically updated. We are also planning the use of Visual Retrieval Techniques for the automatic description of shape, scale and colour distribution of large set of 3D files, as a complementary tool to browse the repository.

A crucial problem of heterogeneous resource collections in Web-based applications, often preserved in different repositories which have adopted different standards and formats, is their management and visualization with a homogeneous interface. In the DesignNet system, metadata are associated to resources inside a RDBMS with a Web interface

appositely created. This allows effective information retrieval and manipulation through exploitation of Java and XML, access to metadata but also to the very same resources. Visualization is supported by previous selection of current available standards (PDF, JPEG, VRLM, MP3, Real, QuickTime), turning raw files into deliverable products to assure their portability on different platforms.

## 3. The prototype's implementation

Within those general workflow and conditions a first prototype has been implemented. The heart of multilayered structure DesignNet Metadata Management System is DesignNet framework, based on a J2EE platform and on a Dublin Core Metadata Schema structure. The framework is a set of tools which enables to create, catalogue and search items recorded within the database. Content is collected, selected and processed with metadata creation and validation. The Industrial Design Thesaurus is both an information storage and retrieval tool. It's used by indexers as a listing of words and phrases authorized for use, showing with relationships, variants and synonyms. For searchers it's a aid to navigation and retrieval, sitting behind the search interface.

An input tool with HTML form allows content providers and project team members to: create new records, search and browse existing ones, verify and give them authority, update them when needed. A RDMBS repository stores the records and a application provides a searchable and browsable interface using a Apache 1.3.9 Web server. The database can support queries based on terms from the Industrial Design Thesaurus, authority files, Dublin Core Elements and Qualifiers. As the framework has been built to be flexible, at any time tables can be added or deleted without having to change the entire schema of the database.

In order to represent semantic values while searching documents, dynamically build schemas provide to help users in the retrieval and to evoke the context of the searched resources, in terms of quantity and typology. The system shows users their search path, stirring up a major consciousness of the researching context and methodology. Moreover, there are some features to trace our activity: tags can be used to insert an object into advanced bookmarks listed by category or related to a real project. Users can see this tag and ask the system such information.

The system features a single point of entry for the user to cross-database digital resources; knowledge-based retrieval with dynamic visualization; advanced search and browse functionalities. Also available are: documentation for DesignNet Team members and users (with glossary, resource selection criteria, metadata manual, progress report, publications); online white-boards for message exchange and e-bulletins for user feedbacks; maps of people competences; sections with on-going projects and latest resources added; conference and events notice board to promote Italian design knowledge; electronic forum for linking students, researcher and professionals for training, long term partnerships, research activities; possibility of sharing files with others on the Internet through upload and download.

*Documentation and references of the project are available at www.designet.polimi.it.*

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 199*


# Metadata hiding tightly binding information to content

Roberto Caldelli, Franco Bartolini, Vito Cappellini
Electronics and Telecommunications Department,
University of Florence,
Florence, Italy

## Summary

The future dramatic development of telecommunications infrastructures (next generation Internet and wide-band mobile networks) will strongly push forward the diffusion of multimedia communication services. Anyway the real effective development of such services will strongly depend on the availability of a reliable platform for managing all the issues related to the exchange of multimedia items among users through heterogeneous networks and systems. Such a need is also witnessed by the ISO MPEG21 initiative whose goal is to achieve "an environment that is capable of supporting the delivery and use of all content types by different categories of users in multiple application domains". In particular some important elements which are considered by MPEG-21 still to be addressed for achieving its goal are Digital Items Identification and Description and Intellectual Property Management and Protection. Some calls have been already issued regarding the identification and description schemes: although it seems that metadata (and XML) will have an important role for addressing this issue, anyway it is evident that much work has still to be done. Future coming *watermarking technologies* will have thus to consider this kind of metadata, and how these will influence their behaviour. It is possible, for example, to suppose that some particular type of metadata should be hidden inside the data themselves for security/confidentiality reasons: these metadata would be known only to those who have knowledge of them (any other person neither would notice their presence) and are authorised to access them. In general this approach would make the embedded metadata independent from the particular format used for storing the image (being this requirement no satisfied if the metadata would have been embedded solely into the image header which is obviously format dependent), and resistant to format changes. In particular metadata embedding is attractive because offer the possibility to make metadata persistent through digital-to-analogue transformations. Of course the need to embed metadata inside the image raises a issue which is beginning to be addressed by watermarking research and regards the technologies to be developed for increasing the watermarking payload given a certain degree of robustness. It is presently emerging with evidence that many of the watermarking techniques developed until now are able to grant payloads that are strongly inferior to what can be theoretical estimated as the capacity limit. These results are encouraging researchers to attempt to design more powerful coding and decoding strategies.

In this paper an application for embedding, inside a digital image, metadata for identifying its IPR status is presented; this insertion has been achieved by means of a digital watermarking technique. This technology has been developed within the IST 21031 Tradex European Project. The metadata have been constructed according to the indications contained in the JPEG standard

**Keywords:** *metadata, Intellectual Property Rights (IPR), digital watermarking, content management.*

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 201-203*


# Learning how to Learn: Using the Dublin Core Metadata Element Set to Support Teachers as Researchers

Patrick Carmichael
University of Reading, UK

## The Teaching and Learning Research Programme

Teachers need to know what they can do in their classroom practice to help pupils acquire the knowledge and skills of learning how to learn. At the same time, the transfer of knowledge among teachers and between networked schools needs to be investigated and an evidence-based model of knowledge creation and transfer in school settings needs to be developed.

The Teaching and Learning Research Project (TLRP) is a £ 23 million programme of research into teaching and learning in the United Kingdom commissioned and managed by the Economic and Social Research Council. The programme aims to "enhance the achievement of learners at all ages and stages in education, training and life-long learning; develop the capability for transforming the knowledge base relevant to learning into effective and efficient teaching and training practices; enhance the system-wide capacity for research based practice in teaching and learning [and] promote and extend multi-disciplinary and multi-sector research in teaching and learning". (Teaching and Learning Research Programme n.d.). Central to the programme's approach is a commitment to use research in support of "evidence-based teaching and learning", characterised by Macintyre and Macintyre (1999, p15) as being "concerned with the effectiveness of patterns of teaching and learning, or with 'what works'". The programme's approach is premised on a view that the improved understanding of educational practices offered by educational research leads to more informed and effective policies and practice through teacher education, the development of curriculum materials and exemplification of 'best practices'. At the same time, educational research is advanced and sustained by the presence of 'teachers as researchers' engaged in action research on their own practice (Stenhouse 1975), and many of the projects which make up the Teaching and Learning Research Programme have a commitment to involving and supporting teachers in research activity.

## Learning how to Learn

'Learning how to Learn' is a project in the second phase of the Teaching and Learning Research Programme and has been running since January 2001. It involves over 40 schools, spread across 6 Local Education Authorities in the UK, in a programme of training and development as part of which they identify areas of potential development of their assessment practice. They are supported in this by an assigned 'critical friend' – a member of the project team who facilitates training and needs analysis, advises on the application of new strategies, and supports teachers who wish to undertake research in their own classrooms.

The project itself builds on previous work: in particular, the work of the Assessment Reform Group (Assessment Reform Group, 1999) and of KMOFAP (Kings-Medway-Oxfordshire Formative Assessment Project) (Black and Wiliam, 2000). These in turn draw on the work of Black and Wiliam (1998a, 1998b) whose review of research into classroom assessment informs both the 'research-based principles to guide classroom practice' of the Assessment Research Group (Assessment Reform Group, 2001) and the approach to the development of classroom practice which underlies the current project. Black and Wiliam (1998b, p. 13) argue that:

"teachers will not take up attractive-sounding ideas, albeit based on extensive research, if these are presented as general principles which leave entirely to them the task of translating them into everyday practice … what they need is a variety of living examples of implementation, by teachers with whom they can identify and from whom they can derive conviction and confidence that they can do better and see concrete examples of what doing better means in practice".

At the same time, the intention is not simply to present teachers with 'recipes' for successful practice, but rather to support them in undertaking research and development in their own classrooms and to explore theoretical insights and research evidence

underpinning the classroom practice, extending and elaborating what Elliot (1991, p. 54) describes as 'a theory of education and teaching which is accessible to other teachers'. This is to be achieved, in part, through access to a developing online 'Knowledge Base'.

The Knowledge Base comprises a collection of resources including text (including accounts of classroom practice, transcripts and children's writing), images, audio and video content. These illustrate practice in a number of areas: 'Questioning' (concerned with effective teacher questions); 'Quality' (concerned with teachers making explicit to learners what measures of achievement they use); 'Feedback' (the nature of teacher response to learner's work); and 'Self-Assessment and Peer Assessment'. They illustrate teaching and learning in different curriculum areas with learners of different ages in a variety of classroom settings. In addition, there is a series of general pedagogical principles derived from the work of KMOFAP and Assessment Reform Group, each of which is supported by research evidence. Metadata records of relevant published and unpublished research reports are also incorporated into the Knowledge Base.

## The Learning how to Learn Metadata Set

While some of the entities within the Knowledge Base are relatively easy to describe as 'learning objects' using Dublin Core (the Qualified Dublin Core Metadata set is used throughout, principally to allow the expression of the frequently rather complex patterns of authorship, editorship and other 'contributor' roles), it has proved necessary to combine it with other metadata sets and to design our own set of elements and qualifiers in order to describe fully all project resources – particularly those which describe in 'fine-grained' detail the classroom strategies and activities which we were presenting to teachers as representing exemplary practice. While substantial numbers of sites across the World Wide Web provide teachers and trainee teachers with ready-made 'lesson plans' and other classroom resources (which can, of course, be described quite adequately using Dublin Core), use of these does not in itself promote good practice in the areas with which the project was concerned.

Resources in the Knowledge Base are, therefore, described using an XML-RDF framework using a combination of elements drawn from the Qualified Dublin Core Metadata Set, the IMC's VCalendar and VCard schemes and our own 'Learning how to Learn' namespace. The decision to implement the Knowledge Base in RDF was informed by a need to express complex relationships between components and draws extensively on Kokkelinck and Schwanzl's (2001) discussion of the implementation of Qualified Dublin Core in RDF.

Where possible, we have used Dublin Core elements so that, in the event of the Knowledge Base subsequently being indexed by a Dublin Core-compliant application, basic information about the resource will be retrieved in accordance with the 'Dumb-down' principle. At the same time, the concern of the project to provide teacher-researchers with suggested classroom strategies and associated exemplars along with pointers to the 'evidence-base' informing their use made it necessary to extend the metadata set used to describe resources. After consideration of existing schemes that extend the Dublin Core such as the GEM (Gateway to Education Materials) metadata set (GEM, 2002), and IEEE Learning Object Metadata element set (IEEE, 2001), a project-specific namespace capable of describing classroom teaching strategies in 'fine grained' detail was developed. This was justified on three grounds:

Firstly, many of the strategies identified by the Assessment Reform Group and by KMOFAP and advanced by the project are designed to be integrated into teachers' existing classroom practice; some involve regular interventions each of only a few minutes' duration and others involve teachers' interactions with individual learners or small groups within the scope of normal classroom activities. We address this need by using a 'description' tag and also use the VCalendar recurrence rule grammar to describe repeated learning activities.

Secondly, we wish to present teachers with clear rationales for the implementation of new practices in assessment, wherever possible related to research evidence, and this requires greater detail than currently offered by the IEEE LOM 'Educational' or the GEM 'Pedagogy' metadata elements. A 'rationale' element is included within the namespace and is used to link exemplars to underpinning project principles.

Thirdly, the 'living examples of implementation' we present to teachers are drawn from a range of classroom contexts, and in many cases are offered as suggestions and stimuli for evaluation and possible action; the notion of 'audience' (as used by many of the educational metadata schemes including Dublin Core; Dublin Core Education Working Group, 2002) is inadequate to describe this purpose. Instead, the project namespace includes a qualified 'context' element which allows the 'origin' of the strategy to be distinguished from its 'application' – other classroom contexts, audiences or curriculum areas in which it has been, or might be, applied.

## Implications and Prospects

The existence of an extended metadata vocabulary capable of describing not only learning resources but also the classroom contexts in which they may be used, the strategies underpinning them and associated research and other publications has allowed us to begin building a sophisticated Knowledge Base not

only capable of addressing Black and Wiliam's (1998, p. 13) call for "living examples of implementation", but also of stimulating teachers to extend the scope of the resources on the basis of their own developing classroom practice. The Knowledge Base architecture will allow web pages to be constructed which offer teachers structured information about classroom activities appropriate to their particular circumstances, together with illustrations (on demand) of their practical implementation. They will be able to comment on the activities and on their experiences of their implementation and offer further illustrative material for integration into the Knowledge Base in order to extend its scope. In addition, they will be able to relate their use of classroom strategies to the broader aims of the project and to school and Local Education Authority priorities, and will be able to locate their practice in a broader theoretical context.

The 'Learning how to Learn' website, which contains further information about the development of the project namespace and its application in the Knowledge Base is located at http://www.learntolearn.ac.uk.

## References

Assessment Reform Group, 1999, *Assessment for Learning: beyond the black box* (University of Cambridge School of Education: Assessment Reform Group).

Assessment Reform Group, 2001 *Assessment for Learning: 10 Principles* (University of Cambridge School of Education: Assessment Reform Group). Available at: http://www.assessment-reform-group.org.uk/principles.html [Accessed 24.06.2002].

Black, P. and Wiliam, D., 1998a. Assessment and Classroom Learning *Assessment in Education* 5(1) p. 7-71.

Black, P. and Wiliam, D., 1998b. *Inside the Black Box: Raising Standards through Classroom Assessment* London: King's College London School of Education.

Black, P. and Wiliam, D. 2000. *The King's Medway Oxford Formative Assessment Project: a theoretical framework for formative assessment?* Paper presented at 'Getting Inside the Black Box: Formative Assessment in Practice' Symposium, British Educational Research Association 26th Annual Conference, Cardiff University, September 2000.

Dublin Core Education Working Group, 2002. *Proposal for audienceLevel Qualifier for the Audience Element*. Available at: http://www.ischool.washington.edu/sasutton/dc-ed/Audience-Level-Proposal.html [Accessed 24.06.2002].

Elliot, J., 1991. *Action Research for Educational Change* (Buckingham, Open University Press).

Gateway to Educational Materials, 2002. *GEM 2.0 Elements and Semantics* Available at: http://www.geminfo.org/Workbench/GEM2_elements.html [Accessed 24.06.2002].

IEEE, 2002. *Draft Standard for Learning Object Metadata* Available at: http://ltsc.ieee.org/wg12/LOM_WD6.doc [Accessed 24.06.2002].

Kokkelinck, S. and Schwanzl, R., 2001 *Expressing Qualified Dublin Core in RDF/XML*. Available at: http://dublincore.org/documents/2001/08/29/dcq-rdf-xml/ [Accessed 24.06.2002].

McIntyre, D. and McIntyre, A. 1999. Capacity for Research into Teaching and Learning: Report to the Teaching and Learning Research Programme. *Unpublished ESRC Report*. Available at: http://www-tlrp.educ.cam.ac.uk/docs/mcintyre.doc [Accessed 27/06/02].

Stenhouse, L. 1975. An Introduction to Curriculum Research and Development, London: Heinemann.

Teaching and Learning Research Programme, n.d *The Teaching and Learning Research Programme: Objectives* Available at: http://www.tlrp.org. [Accessed 24.06.2002].

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 205-206*


# The Need for a Meta-Tag Standard for Audio and Visual Materials

Diana Dale, Ron Rog
Department of Canadian Heritage
Government of Canada
diana_dale@pch.gc.ca and ron_rog@pch.gc.ca

## Abstract

*In Canada, as elsewhere around the world, government is trying to move into the Internet Age, to communicate more and more interactively with an ever-increasing portion of the electorate and to increase the interoperability of digitized media.*

*The Canadian Government Online Initiative, of which we are a part, is an example of this trend.*

*To facilitate access to our materials, we need metatags, metatags that, by and large were originally set up to deal with print media. Thus, we have been struggling in recent years to apply metadata to a test database of Canadian cultural audio and visual clips that we call "Heritage Line".*

*We have followed many avenues for making our data searchable and accessible. We have used the Dublin Core schema, both with the qualified set of elements as well as the unqualified set. Our problems arose specifically with respect to the elements 'type' and 'format'.*

*Mpeg-7 currently appears to offer a solution to our problem.*

## What Is Mpeg-7 and Why Is It Important?

Mpeg-7 is the new standard for multimedia description approved by the International Standards Organization (ISO #15938) in 2001.

It is important because it is a standard aimed at furthering the use of metadata, such as that contained within Dublin Core, in the description and retrieval of audio and video materials. Mpeg-7 is an enabling standard, in that it facilitates the incorporation of any form of metadata, whether proprietary or public, into its structure, in order to expand its capability for searching and sharing data.

The fundamental approach of Mpeg-7 is to describe and search digitized materials by means of sampling, as well as by using lexical search terms. As the amount of data grows, and newer technologies get implemented, the requirements for metadata usage and searching capabilities will grow proportionately.

Understanding the value of the Mpeg-7 standard and being involved at its conception will prove invaluable to any organization trying to send and search audio and video material in a web environment.

In addition, the use of XML would be beneficial for presenting audio and video via various channels of delivery.

## Canadian Cultural Heritage Metadata Project Involving Mpeg-7

The Mpeg-7 Working Group within the Department of Canadian Heritage has three projects currently under way that relate to possible applications of this standard. The group comprises representatives from the Department of Canadian Heritage and such portfolio agencies as the National Archives, the National Film Board of Canada and the Canadian Broadcasting Corporation.

These projects are:

- Running Mpeg-7 tools on audiovisual databases from National Archives, National Film Board, Canadian Heritage and the CBC so that interoperability of materials using disparate metadata schemes (RAD, MEDOC and Dublin Core) can be tested in a web based environment.
- We are also working on demonstrating the search functionality of Mpeg-7 by running its incorporated extractor tools on "Heritage Line", the Canadian Heritage database of audio and video clips.

## The Process

The first step is to collect database samples from working group partners to best represent all the dis-

parate metadata schemes. These samples will then be tagged using MPEG-7 metadata tools and inserted into the index of Heritage Line. The metadata will be mapped from MPEG-7 to Dublin Core elements, and a chart will be created to show the relationship between MPEG-7, Dublin Core and the individual schemas of each database. Finally, search testing will be conducted to verify the accuracy of the tools in mapping and functionality.

## Concluding Remarks

By 2005, the Government of Canada intends to offer all services online. The Dept. of Canadian Heritage, the custodian of the largest inventory of cultural media in the country will be well positioned to index, search and exchange multimedia across the web. The groundwork being laid through working with the Mpeg-7 standard and tools will be the glue that brings together multi-format databases and media. The metadata embedded within the Mpeg-7 standard already maps to the Dublin Core metadata element set. This built-in interoperability will result in the convergence of the library metadata and the multimedia resource communities for the betterment of end users requiring access.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 207-211*


# Enhancing end user searching of HealthInsite

Prue Deacon
HealthInsite Editorial Team
Commonwealth Department of Health and Ageing, Australia
prue.deacon@health.gov.au

## Abstract

*HealthInsite is the Australian federal government's Internet gateway to quality health information. The site was an early adopter of Dublin Core and makes extensive use of metadata in its navigation structure. HealthInsite has two search options utilising the Verity search engine: a simple text search and a metadata search. A third search option is the thesaurus search which is most likely to be used by information specialists. Additional functionality is being considered to improve subject searching for end users. This paper defines the research needed as background to developing the system specifications. The need to consider the whole information retrieval process is emphasised, and a clear role for metadata specialists identified.*
**Keywords:** *Dublin Core, end user searching, HealthInsite, metadata, search engines, subjects, subject element, thesauri*

## Dublin Core metadata and search engines

The main purpose of Dublin Core metadata is to promote relevant resource discovery by enabling more precise searching. However metadata cannot be implemented in isolation; it must be considered as part of an information retrieval system. There is no value in creating metadata if there is no system with the search functionality to utilise it.

In the early days of Dublin Core, there was some expectation that the public search engines would take it up. This might have happened if all implementations had used simple DC with no qualifiers and schemes. In practice, most implementations needed some complexity to give real value. Furthermore, different implementations needed complexity in different areas. Theoretically it is possible to dumb down any DC metadata to simple DC, but this is of limited value and certainly the public search engines have not rushed in to do so.

Thus DC implementations tend to be in relatively closed systems with limited interoperability. Such closed systems are small compared with the whole web and the value of metadata may well be less obvious. A good search engine performing text searching with appropriate ranking will achieve satisfactory results for many searches. A user who moves on to a metadata search may find that it appears to be no better than a text search. The user may even be confused by all the search options offered.

I believe that metadata can add considerable value in a closed/small system but that, to exploit it, you need to go beyond the standard search engine functionality. Metadata developers need to work closely with search analysts and system developers to get the most out of metadata.

In our gateway site, HealthInsite <http://www.healthinsite.gov.au>, we feel that improvements are needed in the user search functionality, particularly for subject searching. This is likely to require some new applications work which could be costly. Before starting we need to do some research into user experiences on HealthInsite, end user behaviour in general and the benefits that could come from different search applications.

## HealthInsite background

HealthInsite is the Australian federal government's Internet gateway to quality health information. The site is managed by the Commonwealth Department of Health and Ageing. HealthInsite works through information partnerships with authoritative website owners ranging from government agencies to private non-profit organisations and support groups. Partners undergo a quality assessment process and then HealthInsite links to individual resources on their sites. Currently there are 54 partners and nearly 9000 resources; 50% of the resources are consumer health information, 30% are written for a health professional/provider audience and 20% are intermediate. HealthInsite also links to international health gateways with similar aims and quality assurance. HealthInsite was launched in April 2000 with a limit-

ed coverage of health topics. This has now been considerably expanded. The next phase is to examine portal functionality, including the provision of access to services.

Our department was an early adopter of Dublin Core, first for the departmental website and then for HealthInsite. Our decision to use DC was in accord with thinking at whole-of-government level in Australia. We have been closely involved with AGLS development (Australian Government Locator Service <http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html>). For us, the advantages of Dublin Core are: simplicity; the delineation of key elements for resource discovery and display; and international and national recognition.

Simplicity is a crucial feature. We have tried to keep as close to simple DC as possible on the grounds that "the simpler the indexing structure, the easier it is to design search applications".

Figure 1 summarises our metadata specification.

Our modus operandi is that information partners maintain metadata on their own sites, usually embedded in the HTML coding of a resource but sometimes located in a separate directory. The metadata is harvested into the HealthInsite Oracle database and harvested again at regular intervals to pick

| Element .qualifier | Scheme | Data format of content | Usefulness* |
|---|---|---|---|
| DC.Creator | | text | metadata group 2 |
| DC.Publisher | | text | metadata group 2; display |
| DC.Rights | | text | partner site administration |
| DC.Title | | text | metadata group 1; display |
| DC.Title.Alternative | | text | metadata group 1 |
| DC.Subject | Health Thesaurus | text terms (controlled vocabulary); semi-colon delimiter between terms | metadata group 1 |
| DC.Description | | text | metadata group 1; display |
| DC.Language | RFC1766 / 3066 | 2-3 character codes; semi-colon delimiter between codes | limit |
| DC.Date.Created | ISO8601 | formatted date | partner site administration |
| DC.Date.Modified | ISO8601 | formatted date | limit; display; personalisation |
| DC.Date.Issued | ISO8601 | formatted date | partner site administration |
| HI.Date.Review | ISO8601 | formatted date | partner site administration |
| HI.Date.Reviewed | ISO8601 | formatted date | partner site administration |
| HI.Date.Healthinsite | ISO8601 | formatted date | personalisation |
| DC.Type | HI type | text term from menu | limit |
| DC.Type | HI category | text terms from menu; semi-colon delimiter between terms | limit |
| DC.Format | IMT | text term from menu | limit |
| DC.Format.Extent | | numeric (size in Kb) | recorded, but not yet used |
| DC.Identifier | URI | URL | link |
| AGLS.Availability | | text | recorded, but not yet used |
| AGLS.Audience | HI age | text term from menu | limit |
| HI.Complexity | | text term from menu | limit |
| HI.Status | | text term from menu | HealthInsite administration |

*Usefulness code:

Metadata group 1: Title, subject, description grouped together in the Healthinsite metadata (power) search.
Metadata group 2: Creator, publisher grouped together in the Healthinsite metadata (power) search.
Limit: can be used to limit a search or for ranking/sorting the search results.
Display: title, description, publisher, date.modified are the elements displayed in search results sets.
Partner site administration: Elements for partners to use, if they wish, in managing their websites.
Link: used to link from the results set to the resource on the partner's site.
Personalisation: used in managing the personalisation features of HealthInsite.

**Figure 1. Summary of HealthInsite metadata specification**

up any changes. In practice, things are a little more complicated. We assist many of our partners with creating the initial metadata records and we create the subject element for most records.

Figure 2 shows the metadata record for one of our partner sites.

## The HealthInsite technical platform and Verity search engine

The technical platform for HealthInsite comprises: an Oracle database for metadata; a modular Cold Fusion application for presentation (soon to be replaced by the Spectra content management system); and a Verity search engine. When we implemented Verity, we were advised that some of the ideas we had for search functionality were beyond the scope of a search engine and had to be deferred for separate development.

As a search engine, our implementation of Verity can index text and index metadata. It enables searches based on full text (simple search) or restricted to text in groups of metadata elements (metadata or power search). It allows Boolean logic, truncation, limiting by various metadata elements and ranking/sorting by various metadata elements. The current implementation does not cater for spelling mistakes and synonyms.

## The subject element in HealthInsite

HealthInsite is a subject gateway and it is known that most searches will be for subjects – the subject element is the focus for the rest of this paper.

Subject indexing in HealthInsite is very tightly controlled. We use the Health Thesaurus <http://www.health.gov.au/thesaurus.htm> which is a hierarchical thesaurus based on MeSH (Medical Subject

```
<META NAME="DC.Creator" CONTENT="Department of Human Services (Victoria)">
<META NAME="DC.Creator" CONTENT="Centre for Eye Research Australia (CERA)">
<META NAME="DC.Publisher" CONTENT="Better Health Channel">
<META NAME="DC.Rights" CONTENT="">
<META NAME="DC.Title" CONTENT="Diabetic retinopathy">
<META NAME="DC.Subject" SCHEME="Health Thesaurus" CONTENT="causes; complications; diabetes
  mellitus; diagnosis; lasers; retinal diseases; risk factors; surgery; symptoms">
<META NAME="DC.Description" CONTENT="Diabetic retinopathy is an eye disease caused by complications
  of diabetes. Everyone with diabetes will develop diabetic retinopathy. Regular eye exams when first diag
  nosed with diabetes, and then at least every two years, will reduce the risk of vision loss and blindness.">
<META NAME="DC.Language" SCHEME="RFC1766" CONTENT="en">
<META NAME="DC.Date.Created" SCHEME="ISO8601" CONTENT="2000-03-08">
<META NAME="DC.Date.Issued" SCHEME="ISO8601" CONTENT="2000-03-20">
<META NAME="DC.Date.Modified" SCHEME="ISO8601" CONTENT="2001-04-12">
<META NAME="DC.Date.ValidTo" SCHEME="ISO8601" CONTENT="">
<META NAME="DC.Date.Review" SCHEME="ISO8601" CONTENT="2002-04-12">
<META NAME="DC.Date.Reviewed" SCHEME="ISO8601" CONTENT="2001-04-12">
<META NAME="DC.Type" SCHEME="HI type" CONTENT="document">
<META NAME="DC.Type" SCHEME="HI category" CONTENT="resource">
<META NAME="DC.Format" SCHEME="IMT" CONTENT="text/html">
<META NAME="DC.Identifier" SCHEME="URI" CONTENT="http://www.betterhealth.vic.gov.au/bhcv2/bhcar
  ticles.nsf/pages/Diabetic_retinopathy">
<META NAME="AGLS.Availability" CONTENT="">
<META NAME="AGLS.Audience" SCHEME="HI age" CONTENT="adult">
<META NAME="HI.Complexity" CONTENT="easy">
<META NAME="HI.Status" CONTENT="registered">
```

Note that on the Better Health Channel site, this resource has an additional subject keyword string: bleeding eye, blindness, Centre for Eye Research Australia, CERA, diabetes, diabetic eye disease, diabetic retinopathy, Diabetes mellitus, Diseases and Disorders, Endocrine Diseases, endocrine, laser treatment, loss of vision, macula, macula vision, maculopathy, proliferative retinopathy, retina, Retinal diseases, Eye Diseases, sightless, vision, vision loss.

**Figure 2. Metadata from HealthInsite for a resource on the Better Health Channel, a HealthInsite information partner - HTML syntax**

Headings) <http://www.nlm.nih.gov/mesh/meshhome.html>. Indexing is as specific as possible using preferred terms from this thesaurus. In the metadata record, the subject element looks quite simple. For example, from Figure 2:

<META NAME="DC.Subject" SCHEME="Health Thesaurus" CONTENT="causes; complications; diabetes mellitus; diagnosis; lasers; retinal diseases; risk factors; surgery; symptoms">

This subject line provides some useful words for resource discovery in the metadata search option. However, there are more sophisticated search possibilities. When the subject string is pulled into HealthInsite, the subject terms are associated with their hierarchy numbers. For example, diabetes mellitus has the numbers C.018.452.297 and C.019.246. This relates it to the broader terms "metabolic diseases", "nutritional and metabolic diseases" and "endocrine diseases" in the disease schedules of the hierarchy. It also relates it to the narrower terms "insulin-dependent diabetes mellitus" and "non-insulin-dependent diabetes mellitus"

Expert searchers, with knowledge of the thesaurus hierarchy and Verity, can use the full power of the thesaurus when searching. They can use the hierarchy as well as the related term structure to perform complete, but precise, searches. For example, in the HealthInsite navigation/browse facility, which is a topic-based structure, each topic contains an expert search which is performed dynamically on the latest version of the database.

For example, the topic "Drug treatments for heart disease" has the search

( c.014.280* <IN> THESAURUS_TREE_CODE or cardiology <IN> THESAURUS_TERM_NAME ) and e.002.319* <IN> THESAURUS_TREE_CODE

In this search c.014.280* picks up "heart diseases" and all its narrower terms; e.002.319* picks up "drug therapy" and all its narrower terms.

This topic query technique is a major feature of HealthInsite, enabling considerable flexibility in adjusting topics without having to adjust metadata. It was evaluated in an earlier collaborative study (Deacon, Buckley Smith & Tow, 2001). These complex searches are clearly not an option for end users.

Currently HealthInsite has a thesaurus search option which allows users to navigate up and down the hierarchy (one step at a time) or to search on preferred terms. The interface is relatively limited, not self explanatory and may confuse the user. With some terms, the user would be much better to do a simple text search.

For example, a text search on nappy rash leads to 19 documents, of which the first 5 are highly relevant and the rest might have some useful information. It would take users 3 steps to get to the thesaurus search page. There they would find that nappy rash is not a valid thesaurus term. They would then have to work out what to do next.

In contrast to HealthInsite, one of its information partners, the Better Health Channel <http://www.betterhealth.vic.gov.au>, uses a controlled keyword scheme for subjects. In the metadata record in Figure 2, the keyword string is:

"bleeding eye, blindness, Centre for Eye Research Australia, CERA, diabetes, diabetic eye disease, diabetic retinopathy, Diabetes mellitus, Diseases and Disorders, Endocrine Diseases, endocrine, laser treatment, loss of vision, macula, macula vision, maculopathy, proliferative retinopathy, retina, Retinal diseases, Eye Diseases, sightless, vision, vision loss"

This has far more handles for resource discovery by an end user than the subject element in HealthInsite. The end user probably would not notice that some types of searches in the Better Health Channel site would lack precision.

In summary, while the metadata subject framework is essential for the topic-based navigation facility on HealthInsite, it is of limited use to end users doing their own searches.

## What improvements could we make?

We feel that that the current situation is unsatisfactory for end users and that subject search functionality should be improved. These are some of the options:

- Bring the full librarians' functionality into an end user framework – like the subscription versions of Medline <http://www.nlm.nih.gov/databases/databases_medline.html>, or the public version (PubMed) <http://pubmed.gov>.
- Provide automatic synonym searching and spell checking.
- Make a link to the thesaurus application and add an application to help users construct searches. (The thesaurus is a database and there is an in-house application which enables full searching, with links between the preferred terms and hierarchy).
- Create standard limits to help users with text searches that retrieve very large results sets – for example, if a user searches on diabetes they could get the option to limit their search to prevention of diabetes.
- Create standard hedges to help people broaden a search. For example, a hedge for "heart" would contain the heart anatomy terms, all the heart diseases and cardiac surgery.
- Enhance the link between user searches and the relevant HealthInsite topics.
- Offer a librarian search service.
- Do nothing – it may be that end users do not really have a problem. If users get some information that

they need, then it may not really matter to them if they have not found all the relevant items or if the results set is not very precise.

Most options require applications development or purchase, some at high cost. Because of the cost, we need to be very clear what we are trying to achieve and that it has real value before writing specifications.

## Research required

The research plan is to study end user subject searching behaviour (both in general and on HealthInsite), to identify where users may require assistance on HealthInsite and to describe what sort of search functionality could provide this assistance.

It is well known that most users will try a simple text search first and many will not try anything more complicated. A literature search is needed, particularly to find evidence on user reactions to advanced or metadata searching.

From HealthInsite, we have three sources of information on end user searching:

- The data files of actual searches performed. These show the type of search (simple, metadata or thesaurus), the number of times the search was performed within a particular period and whether the search was successful or not. With around 2000 visitors a day to HealthInsite, these files are very large.
- User feedback on the site – users may advise us if they have had trouble trying to find information on a particular subject.
- Feedback from focus groups on the sort of facilities users want on HealthInsite. Consumer consultation is an important mechanism within the broader HealthInsite strategic planning process. Specific queries relating to end user searching and usability testing could be incorporated in the next rounds of consultation.

The main research task is to sample user searches from the data files, try the search on all three options (simple, metadata and thesaurus) and then evaluate the success of the search (recall/precision analysis) against the difficulty of performing it.

This will lead to reviewing the unsuccessful searches (including those identified in user feedback) to see what sort of assistance could be given and at what point.

Next, close liaison is required between content managers (metadata and search specialists) and IT staff to identify possible search functionality and its usability. This would involve looking at the options suggested in the previous section above. It will be necessary to assess whether the new functionality is convenient enough for the user to be persuaded to take the extra step beyond a simple text search. Furthermore, if a simple search is satisfactory, would the user be worse off by trying the new functionality?

It will be useful to review search options on other sites, although it is not always easy to ascertain the algorithms used.

There may be implications for the metadata specification or indexing rules – it is possible that a minor change to the metadata could have considerable benefits. Also, there may be new ways to use some of the other metadata elements to enhance subject searches.

## Conclusion

This paper describes the metadata used in HealthInsite and shows that the subject element currently has more value for experts than for end users. The research planning phase of a project to improve subject searching for end users is outlined. When this research is complete, we will be able to decide what is feasible within our technical budget and then prepare the specifications for new search functionality. It is clear that this sort of system enhancement needs to be cognizant of the whole information retrieval process. All players should be involved – metadata, search & navigation and IT specialists, through to end users. The metadata experts in particular have a clear role to ensure the best use of metadata as well as to be flexible in considering adaptations to metadata standards.

## References

Deacon, P., Buckley Smith, J. and Tow, S., 2001. Using metadata to create navigation paths in the HealthInsite Internet gateway. *Health Information and Libraries Journal*, 18, 20-29.

## Web sites and resources

AGLS (Australian Government Locator Service). http://www.naa.gov.au/recordkeeping/gov_online/agls/summary.html

Better Health Channel. http://www.betterhealth.vic.gov.au

The Health and aged care thesaurus. 5th edition. Commonwealth Department of Health and Aged Care 2001. (short title: The health thesaurus) http://www.health.gov.au/thesaurus.htm

HealthInsite. http://www.healthinsite.gov.au

MeSH (Medical subject headings) http://www.nlm.nih.gov/mesh/meshhome.html

Medline http://www.nlm.nih.gov/databases/databases_medline.html

PubMed. http://pubmed.gov

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 213-215*


# Abstraction versus Implementation: Issues in Formalizing the NIEHS Application Profile

Corey A Harper
Knight Library, University of Oregon
harpc@ils.unc.edu

Jane Greenberg
University of North Carolina at Chapel Hill
janeg@ils.unc.edu

W. Davenport Robertson, Ellen Leadem
National Institute of Environmental Health Sciences
{robert11, leadem} @niehs.nih.gov

## Version 1 of the National Institute of Environmental Health Sciences Metadata Schema

The National Institute of Environment Health Sciences (NIEHS), is an Institute of the National Institutes of Health (NIH), which is a component of the U.S. Department of Health and Human Services (DHHS). As with many governmental organizations, the NIEHS website contains a rich and growing body of important resources for both employees and the general public. The NIEHS library has spearheaded an organization-wide metadata project to provide better access to these resources. Dublin Core was elected for the NIEHS metadata project because it supports semantic interoperability and the potential for data sharing, and because the schema is simple enough to support author-generated metadata. This paper and corresponding poster document issues in formalizing the NIEHS Application Profile, specifically the changes implemented between Version 1 to Version 2, which were influenced by revisions made to the NIEHS metadata form supporting author-generated metadata.

Version 1 was comprised of twenty metadata elements, the majority of which were drawn from the basic Dublin Core Metadata Element Set version 1.1, and the expanded set of Dublin Core Qualifiers (Robertson et al. 2001). NIEHS' Version 1 corresponded quite closely with the Dublin Core semantics. Exceptions included merging the *Creator* element with the *Contributor* element, incorporating an *Audience* element from the Gateway to Educational Materials (GEM) namespace. Version 1 of the NIEHS Metadata Schema was an application profile, in the rough sense of the word, but discrepancies in the formal schema and the public schema, which supported the NIEHS metadata form for author metadata creation, delayed official formalization, until version 2 of this schema.

## Version 2: The NIEHS Metadata Schema Becomes an Application Profile

The movement towards an application profile involved resolving discrepancies between the NIEHS' formal schema and the public schema made accessible via the metadata template, identifying the namespaces that elements and qualifiers are derived from, and establishing rules concerning the usage of elements such as the obligation (whether an element's inclusion is required) and cardinality (repeatability of a given element). The most significant questions addresses in defining and documenting a schema were: 1) How many elements are needed? 2) which documents should define element usage rules? and 3) How the documents should be serialized? The following sections correspond to these three issues.

## One Element or Two?

The first main step to revising and formalizing the NIEHS application profile was to determine how many metadata elements were needed for a document attribute. Version 1 of the NIEHS metadata schema identified separate elements for *Date Created*

and *Date modified*, as well as for *URL*, *NIEHS Number* and *Other Identifier*; *Alternative Title* was listed as a separate element despite being a standard Dublin Core Qualifier. These six elements could be condensed to *Date*, *Identifier* and *Title* by applying appropriate refinement qualifiers, as was done with *Coverage*, which was refined by *Time* or *Date* qualifiers.

In Version 2 the six elements highlighted above were reduced to four corresponding Dublin Core elements: *Date*, *Coverage*, *Identifier* and *Title* refined by qualifiers. Many of the qualifiers used are defined in the Dublin Core Qualified (DCQ) recommendation. *URL*, *NIEHS Number* and *Other Identifier* are unique to the NIEHS metadata project, although applied to an official Dublin Core element.

## Namespace versus Application Profile

The Dublin Core Namespace policy, by defining the Dublin Core Metadata Initiative (DCMI) 'Terms' namespace as containing "DCMI elements and DCMI qualifiers (other than those elements defined in the Dublin Core Metadata Element Set [DCMES], Version 1.1)", sets a precedent for defining new elements *and* qualifiers in a single namespace document (Powell and Wagner, 2001, p. 2). This provided justification for the inclusion of any element, refinement qualifier or encoding scheme *that is completely unique to the NIEHS metadata project* in the NIEHS namespace.

After the determination of element representation (one element or two), the NIEHS metadata team determined that *Author/Contributor* was the only element unique enough to warrant a new element definition. Rather than defining a new element combining the Relation and Source elements, the NIEHS namespace defines two refinement qualifiers ('NIEHS Is original source of' and 'NIEHS Has original source') that function like those defined in the DC Terms namespace. A similar issue arose when determining which Encoding Scheme Qualifiers were unique enough to include in the NIEHS namespace. All the value lists used by the project, except the list applied to the Audience element, were derived from existing value lists. It was decided that altered versions of existing lists should also be defined in the NIEHS namespace document.

## Serialization

The third major issue encountered involved the serialization of an element set. XML (Extensible Markup Language) serializations of an element set are useful for providing a machine-readable representation for a local context, such as a template for author generation of metadata. RDF (Resource Description Framework) serializations support semantic interoperability by providing explicit standards for combining varied element sets.

The DCMI registries working group identifies providing "access to DCMI approved domain specific 'application profiles' e.g. the DCMI Education group application profile" as a high priority for the next phase of the DC registry (Heery, 2001, p. 8). In it's current state, this registry contains Resource Description Framework Schemas (RDFS) that define the components of the various DC namespaces. In adherence with this practice, the NIEHS metadata team will encode its application profiles using RDFS. This will ensure that element sets are serialized consistently across initiatives interested in interoperability with NIEHS and vice/versa.

## Conclusions and Future Work

Formalizing the NIEHS application profile was both an intellectual and practical undertaking. As part of the process we discovered that elements incorporated into an application profile can be modified in at least four distinct ways: 1) algorithms can automatically supply metadata (e.g., assign default values or extract information from HTML or XML source code), 2) element designations in a formal schema can be modified to facilitate author and searcher (non-expert) understanding in a metadata template or search engine; 3) cardinality and obligation constraints can be provided; and 4) new qualifiers, or qualifiers from alternate element sets can be applied to existing elements in unique ways. We advocate defining new elements and qualifiers in namespaces modeled on RDF.

Among one of the most significant challenges the NIEHS metadata team needs to address now is how to best codify element cardinality and obligation. RDF *does not* provide any mechanism for this need; and although locally defined XML DTD (document type definition) permit documentation, the local nature provides a significant barrier to interoperability. DAML+OIL (DARPA Agent Markup Language/ Ontology Inference Layer) Reference Description (http://www.w3.org/TR/daml+oil-reference), which is evolving into OWL (Web Ontology Language), provides a standardized mechanism for declaring cardinality constraints. Additionally, these languages permit a layering on top of RDFS and adhere to the dumb-down principle used by Qualified Dublin Core. Query engines and processors 'understanding' DAML+OIL, can extract constraints and understand the intended meaning of attributes from a representation, otherwise, they can pass over and interpret the remainder of the representation as long as the RDF syntax is intact. Here in lies a topic of inquiry and experimentation from the next phase of the NIEHE metadata project. In closing, the topics addressed in this paper can help to better inform the future development of application profiles and name-

spaces, working toward an interoperable environment, one that supports the growth of the Semantic Web.

## References

Heery, R. 2001, Draft DCMI Open Registry Functional Requirements. Dublin Core Metadata Initiative. [Online} Available at: http://dublincore.org/groups/registry/fun_req_ph1-20011031.shtml

Powell, A and Wagner, H. 2001, Namespace Policy for the Dublin Core Metadata Initiative (DCMI). Dublin Core Metadata Initiative. [Online] Available at: http://dublincore.org/documents/2001/10/26/dcmi-namespace/

Robertson, W.D., Leadem, E.M., Dube, J. and Greenberg, J. 2001, Design and Implementation of the National Institute of Environmental Health Sciences Dublin Core Metadata Schema. In: Proceedings of the International Conference on Dublin Core and Metadata Applications 2001. Tokyo, Japan. National Institute of Informatics. p. 193-199. [Online] Available at: http://www.nii.ac.jp/dc2001/proceedings/product/paper-29.pdf

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 217-223*


# Integrating Learning Objects into Learning Contexts

I.T. Hawryszkiewycz
Faculty of Information Technology
University of Technology, Sydney
igorh@it.uts.edu.au

## Abstract

*The paper describes learning objects in the context of a learning process. It examines options of integrating learning objects into context and supporting the integration with learning activities. The paper then examines the technology needed to support the creation and utilization of learning objects. It suggests customizable portals as the solution. It then illustrates an application to teaching.*
**Keywords:** *Portals, Knowledge Management, Learning Objects, Customization*

## 1. Introduction

Learning communities are now beginning to take many forms. There are the conventional classroom situations that still predominate, but increasingly we are beginning to see new forms such as work based learning, distance learning, and virtual universities. Although the learning contexts are different, the material taught can often be based on the same subject material. Increasingly web based technologies are being used to provide services that support these learning environments. Considerable work has taken place in using a variety of such services. Wade and Power [10] for example outlined a number of requirements for computer supported learning systems and described alternate technologies for supporting learning activities. Neal [7] has carried out work on their use in distance teaching emphasizing the delivery of materials. It is however fair to say that much of this research has been in specific settings. Two issues that have been raised as important here are the reuse of learning material in different settings and provision of services through interfaces that are intuitive for learning.

A body of opinion is beginning to form that what is needed, especially for reuse, are learning objects that can be adapted to any of the learning environments. Standards are now being developed for learning objects. Perhaps the two most quoted standards are the Dublin core (http://www.dublincore.org) and the Learning Technology Standards of the IEEE (http://ltsc.ieee.org). These standards describe the elements that are used to describe learning objects this enabling access to these objects to be shared across the WWW. Most learning takes place in a context. This context may be a University or it may be a business entity. Learning objects take a new meaning in their context and can better add to knowledge if they are placed in a context. This paper examines the idea of learning objects and ways to deliver them in context. Context, however, is often specific to the learning environment such as a University or business enterprise. The question arises then on how to combine standard learning objects into the learning context. Two options appear possible here. One to include context in the metadata definition of the learning object, or at least include elements to link to a context. The other is to provide higher level services that integrate the standard learning object into the context using the delivery infrastructure and includes the services needed to support the learning process. The difference is that in the former experiences can be shared by all users of an object, whereas in the latter they are confined to participants in the context.

Another important issue is support for a learning process. The learning object can thus be related to the other dimensions shown in Figure 1. Any learning object is then embedded in metadata and can be linked to other learning objects within the metadata as shown in Figure 1 to facilitate discovery. It is related in a context to provide a goal for learning, and to a learning process to achieve the goal in a most effective manner.

The paper thus examines the elements needed to describe a context from a learning process perspective and then looks at the way this can be integrated with standard data elements. This paper uses Nonaka's knowledge creation process (1994) as underlying theory to define learning on the assumption that learning creates new knowledge either for individuals or groups. The paper uses Nonaka process as a basis for defining learning activities and processes for them. These activities include socializa-

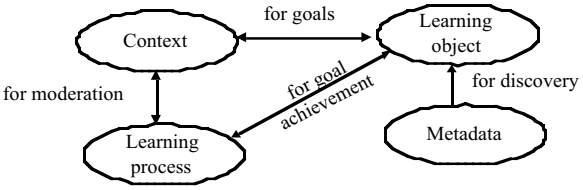



**Figure 1. Objects in context**

tion, developing an understanding of concepts, articulation of ideas, followed by artifact construction and evaluation.

Information technology must then provide ways to create a learning place or environment by integrating standard learning objects into a context and providing activities to support the learning process. From the technological perspective the paper proposes that customizable knowledge portals can be used to integrate learning objects into a context. These resources can include standardized learning objects together with the services and background that make up the learning context. The paper then describes a system, called LiveNet, which can be used to customize such learning models, and ways that it has been used in a teaching environment with a variety of services.

## 2. Evolution Towards Learning Objects

The evolution of learning objects is basically illustrated in Figure 2. Here information is gradually focused through appropriate classification schemes on a particular learning objective and then used to create the learning object. The learning object in many library based system is often restricted to subject material, which must eventually be placed in its context by the learner.

There are in fact two contexts here as illustrated in Figure 3. One is the context within which learning takes place and sets the objective for learning. This outer context may be a University, or a workplace, or a project. The other context is the subject context within which the subject is being taught. This sets a framework for discovery and is usually implemented as links within the metadata structure. Thus teaching for example about databases may place it within the context of businesses or applications. The context of the learning object can also be related to other concepts such as for example how does database design relate to the development process.

## 3. Some Underlying Ideas

Learning can itself be defined as a process that can include a number of roles with responsibilities for maintaining knowledge and passing on their expertise. Such communities are often called as communities of practice.

### 3.1. Communities of Practice

Communities of practice depend on the kind of application. The community of practice can include a variety of roles. The simplest is where there are simply teachers and learners. These can be expanded to include tutors or assistants that work together with the teacher. In more elaborate environments, there can be owners, experts, novices or apprentices as well as a variety of users. They can also be people responsible for specific business process steps. These become the portal roles, each with their responsibilities and provided with appropriate services. Thus the responsibility of the owners may be to create and update the body of knowledge. They can also give permissions to users to access the portals. They can also consult with experts on adding to the body of knowledge. Communities of practice can also include a variety of experts such as subject specialists to discover, classify and distribute knowledge. The IEEE standard defines a variety of roles for this purpose.

### 3.2. Learning Process

Our in defining a learning process is to develop a framework for generic services using the work of Nonaka (1994) as grounded theory. Nonaka sees knowledge sharing and creation following the process shown in Figure 4. These identify the kind of activities that are fundamental to knowledge management.

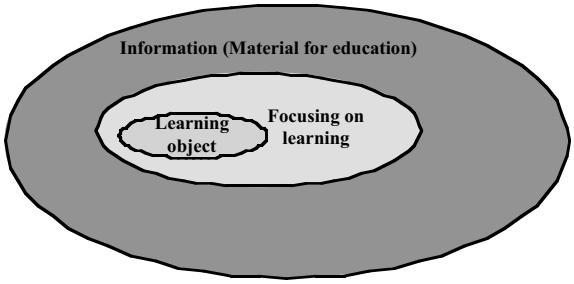



**Figure 2. From information to learning object**

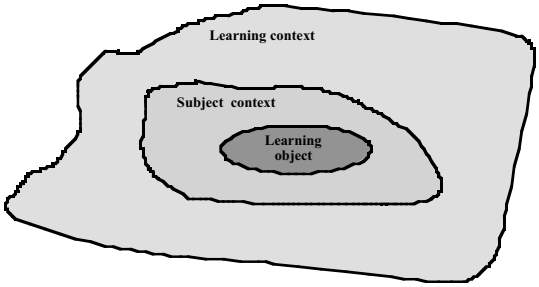



**Figure 3. The Context**

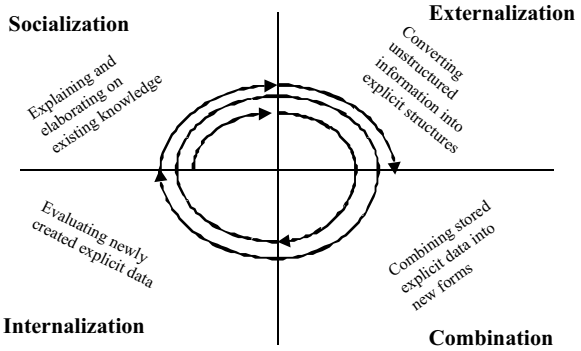



**Figure 4. Nonaka's knowledge creation process**

Nonaka's process includes four phases. The first phase is socialization where people bring together their experiences and share insights in an area. For example, this may be exchange of views of how a particular product was sold to clients. The next step, externalization, is where some of this captured expertise is interpreted into a form that can lead to some actions. In other words how to market the product in a customer's specific context to maximize the likelyhood of achieving a sale. The discussions now focus on identifying new ideas that are now externalized in familiar contexts to see their relevance to specific issues. This often requires the interpretation of new concepts in local terms requiring a clear terminology to articulate the ideas within new contexts. It also includes showing in explicit terms how a product could be used. The ideas are then combined where necessary with existing information and then applied in practice during internalization. Any outcomes of any actions evaluated in further socialization and the cycle is repeated. Nonaka goes further and defines the environments under which knowledge sharing can effectively take place. He suggests that knowledge is only meaningful within a context and its environment. The context defines the relevance of what is discussed and provides the basis for any interpretations. Nonaka defines four different kinds of environments to match his process.

These are:

- Socializing – requires easy and usually informal ways to exchange experiences, develop trust, share values,
- Dialoging – sharing of mental models, articulation of concepts, development of common terms. Usually consciously constructed requiring the interpretation of experiences into familiar contexts,
- Systemising – requires ways to visualize interactions, construct artifacts, combine explicit knowledge and explain how knowledge gained from known experiences is to be used in new ways,

Exercising - communicate artifacts and embody in working context. Reflect on the outcomes.

Our goal is for portals to provide such generic services and provide ways to customise them to particular application needs.

## 4. Learning Structures

Standards are now emerging for learning objects. These generally center on providing ways to classify objects, which in turn is based on an accepted ontology. Learning objects exist within a context and as such should embrace both the context and the body of knowledge. We thus distinguish between a standard for learning objects and a standard for the learning environment. The distinction is illustrated in Figure 5. It shows the learning environment composed of three main parts, namely, the subject material, the context and the learning activities. The latter are defined here from Nonaka's model.

The paper further argues that it is not possible to have a single structure for learning objects but a classification. In that case composite learning objects can be created from more basic objects.

Customization then includes:

- Providing ways to combine the standard subject into the context, and
- Choosing the activities suitable for the learning process.

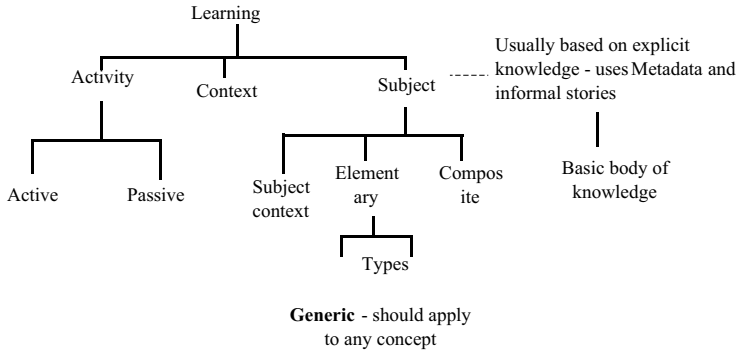



**Figure 5. Classification**

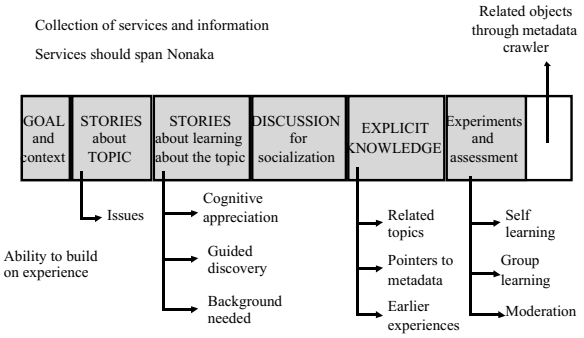



**Figure 6. Generic Structure**

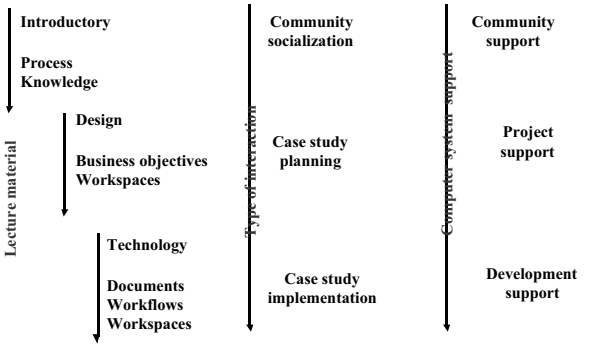



**Figure 8. The learning Process**

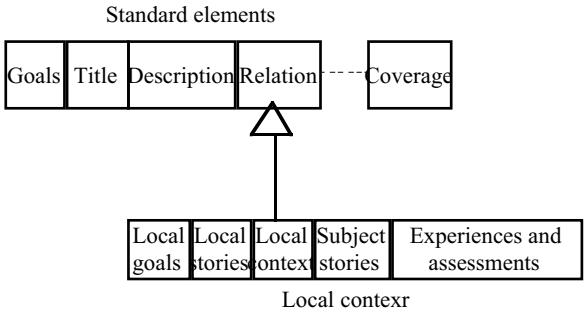



**Figure 7. Integration**

### 4.1. The Learning Infrastructure

The abstract object structure proposed for learning is illustrated in Figure 6. This is the structure that is seen by the learner. It combines Nonaka's framework and contains components that support the aspects of Nonaka's process. These for example, are stories, discussions for socialization, and experiments and assessments for learning within the environment context. The explicit knowledge is predominantly derived from standard learning objects.

Deriving the learning structure from standards is shown in Figure 7. It uses the idea of object inheritance where local learning objects inherit features of standards and enhance them with local content.

## 5. An Example

An example of a subject that uses both approaches is the introduction of technology in its application to electronic business. The way that the subject is taught is illustrated in Figure 8.

- First there is the learning of process and design concepts and ways to describe what business processes. It requires students to understand the design process and its techniques through theoretical exercises. The service here includes a process description and access to exercises and solutions. Socialization is supported to follow-up with questions on the solutions.
- Then various technologies are described. Students here are required to carry out in-depth research in selected topics and provide in-depth but short reports. This requires searches through a variety of objects. Services needed are discovery services and support for providing in-focus documents.
- The students carry out a group case study implementing a system using the methodology. Support is needed here for group interaction and managing of their projects. The services here are to provide group support for joint case study planning and system development.

The concept learning takes place as individuals whereas in the design process students are organized into groups to discuss design alternatives and make design choices. Metadata ideas are useful here to facilitate discovery in both the technology studies as well as the design process. Technology use evolves to support this approach. Initially access concentrates on getting information and socializing. Then a project space is created for each group where alternatives can be considered and design documents maintained. Finally there is the prototype development where students choose technology to implement the design.

The goal is for learners to progress from simple learning of concepts to the application of these concepts in the work environment. It introduces technology and learning in a gradual way. First there is some objectivist learning to describe what business processes using community workspaces. The next step is when the actual design process is introduced and students organized into groups to discuss design alternatives and make design choices. Correspondingly a project space is created in which such alternatives can be considered. Finally there is the prototype development where students choose technology to implement the design.

### 5.1. An example of a metadata structure

We have developed a simple ontology to describe the concepts taught in this subject. These allow learners to create an ontology of related terms and

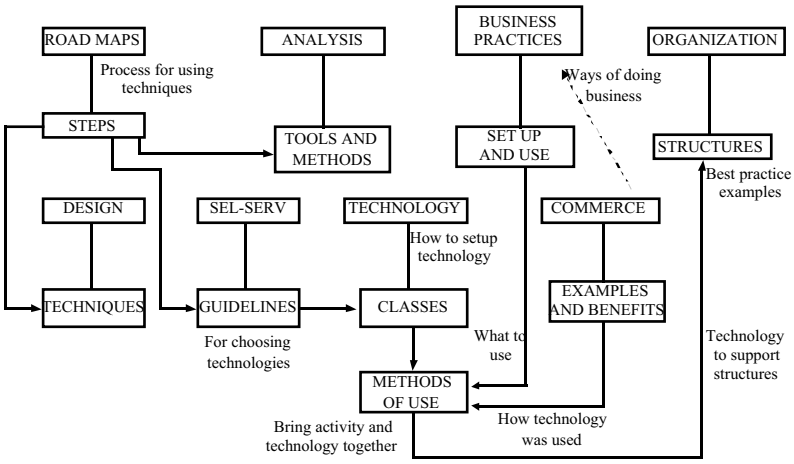



**Figure 9. Structuring the Knowledge**

add to the ontology by elaborating it using insights gained from experience and outcomes in business actions. As an example, we have, developed an ontology for teaching about electronic commerce. This is illustrated in Figure 9. It divides knowledge into seven categories:

- Business practices used in electronic commerce including customer relationship management, supply chains and so on,
- Analysis to describe ways to analyze new systems and define requirements,
- Design approaches to design new systems,
- Commercial applications, business services and
- Technologies used in electronic commerce,
- Business services and how to select technologies to provide them,
- Organizational relationships needed within electronic commerce.

The body of knowledge then contains relationships between these areas. A learner can begin at one concept and then follow links to see how the concept fits into the wider context. Thus it is possible to start with a business practice and then follow links to technology to see what technology can be used to support the practice.

Apart from the ontology of concepts the body of knowledge also includes exercises and solutions, exams, case studies and other study material. It can include previous experiences and suggested actions in a business process step. It can also include guidelines for filling in forms and check-lists for deciding on actions.

## 6. Using Portals For Integration

There are now many portals that make generic services available to users but require the users themselves to choose the most appropriate service for a given business problem. Our goal is to provide was to customize and integrate the generic services for particular business applications. Business services are constructed from the generic services. We illustrate the integration of services needed in the subject described above within our portal.

### 6.1 An Example Portal

Currently we have been using a system, LiveNet, to integrate teaching services. The approach is to emphasize collaboration through an entry interface that emphasizes collaboration while providing access to the body of knowledge. Figure 10 illustrates the basic structure of this interface. It includes menus for defining a community of practice in terms of its roles, interactions between them. The interface shows all the information in the subject. It also provides different roles with different views. Thus for example the folder names 'information-to-tutors' can only be seen by tutors thus reducing the need for meetings and saving peoples time. The interface can then be used to enter the body of knowledge and use its associated knowledge services.

It also provides awareness and notification features to alert members to events important to them. It defines the explicit body of knowledge and providing the actions needed to use it. These include links between objects as well as self-learning through multi-choice questions.

We are currently developing further services to support group formation. Students can form project groups, integrate their learning concepts into the project space and develop a collaborative application. A proposed interface for this purpose is shown in Figure 11.

Here students can form groups, setup meetings, raise issues within the context of a case study. We have used an earlier version of this system concentrating on document management but found that group learning must provide flexible ways to arrange meetings and keep track of progress. The goal here is

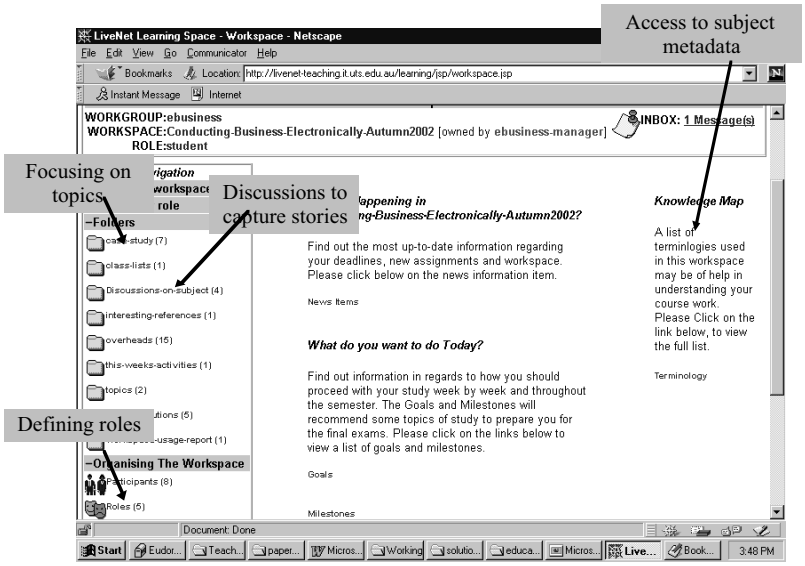



**Figure 10. A LiveNet collaborative services interface**

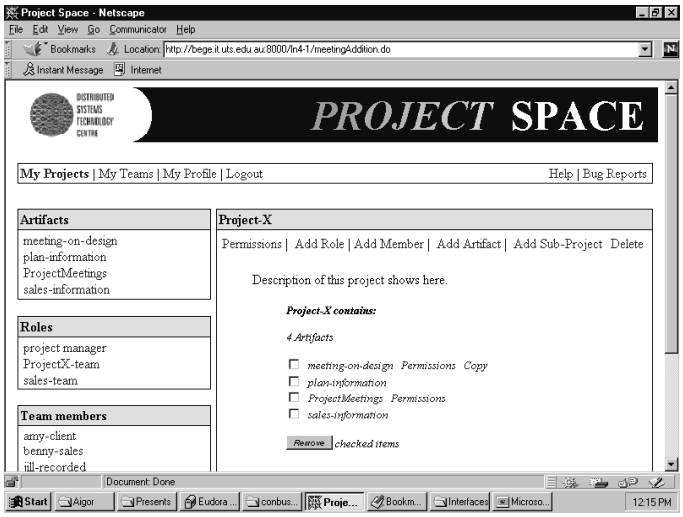



**Figure 11. A Project Interface**

to bring together case study materials, design guidelines and design documents and provide a governance structure, through roles, and facilitate learning through interaction and moderation by teachers, as suggested in the IEEE standard. The interface can include any number of folders, some keeping stories, other concentrating on issues and still others on managing outcomes and collecting suggestions. The structure of these can be customized to the preference of the learners.

### 6.2. Some comments

The learning strategy shown in Figure 8 proved successful in that students apart from various technical problems found the learning of value. This basically introduces technology in gradual stages. These begin with familiarization using the community interface in Figure 9, going on to the private group workspaces for developing project goals and finally through students using the software to develop the prototype for a case study. In the case study students were given a number of milestones to aim for, starting with analysis, through design specification to setting up a prototype LiveNet system. Generally, these were successful in the sense that students understood the basic LiveNet modeling method and workspace description and set up prototypes with little effort. The social effect of this is to require students to pace their work according to the process rather, as is often the case, leaving it to the last minute. This has an obvious learning benefit although it is perceived as a nuisance by some students in that it requires them to follow a process.

## Summary

The paper described the integration of learning objects into their environment through portals. These included of a body of knowledge as well as ways to present knowledge from different perspectives. It also described ways to encourage group learning through flexible project interfaces. Our goal is to determine a range of generic services that should be provided by portals to support learning processes.

## References

Dublin Element Set – http:/www.dublincore.org

Fisher, S. 2001. "Course and Exercise Sequencing Using Metadata in Adaptive Hypermedia Learning Systems" ACM Journal of Educational Resources in Computing" Vol. 1, No. 1, Spring 2001.

Murphy, L.D. 1998. "Digital document metadata in organizations: Roles, analytical approaches, and future research directions" Proceedings of the Thirty-First Hawaiian Conference on System Sciences, Hawaii, 1998, pp. 267-276.

Grant, R.M. 1996. "Prospering in Dynamically-competitive Environments: Organizational Capability as Knowledge Integration" *Organization Science*, Vol. 7, No. 4, July 1996, pp. 375-387.

Hansen, M.T., Nohria, N. and Tierney, T. 1999. "Whats your Strategy for Managing Knowledge" Harvard Business Review, March-April, 1999, pp. 106-116.

Hawryszkiewycz, I.T. 2000. "Knowledge Networks in Administrative Systems" *Working Conference on Advances in Electronic Government*, Zarazoga, Spain, February 2000, pp. 59-75.

Hiltz, R. and Turoff, M. 2002. "What makes learning networks effective?" Communications of the ACM, Vol. 45, No. 4, April, 2002, pp. 56-59.
IEEE Learning Technology Standards Committee – http://ltsc.ieee.org

Jones, C.T., Hesterly, W.S., and S.P. Borgatti 1997. A General Theory of Network Governance: Exchange Conditions and Social Mechanisms. *Academy of Management Review*, Vol. 22, No. 4, October, 1997, pp. 911-945.

Kalakota, R. and Robinson, M. 1999. "e-Business: Roadmap for Success" Addison-Wesley, 1999.

Kuczmarski, T. D. 1997. "*Innovation: Leadership Strategies for the Competitive Edge*" NTC Business Books, Lincolnwood, Illinois.
Leidner, D.E. and Jarvenpaa, S. 1995. "The Information Age confronts education: A theoretical view" Information Systems Research, 4(1), pp. 24-54.

Neal, L. 1997. "Virtual Classrooms and Communities" Group 97, Phoenix, Arizona, pp. 81-90.

Nonaka, I. 1994. "A Dynamic Theory of Organizational Knowledge Creation" *Organization Science*, Vol. 5, No. 1, February 1994, pp. 14-37.
LiveNet – http://linus.socs.uts.edu.au/~igorh/workspace/explore/livenet.htm

Riggins, F.J. and Rhee , H-K. 1998. "Developing the Learning Network Using Extranets" *Proceedings of the Thirty-First Hawaiian Conference on Systems Sciences*, January 1998.

Salmon, G. 2000. "E-Moderating: The Key to Teaching and Learning Online" Stylus Publishing, Sterling, VA.

Wade, V.P. and Power, C. 1998. "Evaluating the Design and Delivery of WWW Based Educational Environments and Courseware" Proceedings of the 6th. Annual Conference on the Teaching of Computing, August 1998, Ireland, pp. 243-248.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 225-227*


# Metadata associated Network Services and Capabilities

Masatoshi Kawarasaki, Junichi Kishigami
NTT Service Integration Laboratories
Kawarasaki.masatoshi@lab.ntt.co.jp, Jay.@ntt.net

## Abstract

*This paper discusses the problems of content delivery in heterogeneous networking environment, and proposes framework architecture based on metadata. The proposed architecture will provide high-quality and user-friendly network services by using metadata about content as well as user's requirement.*

## 1. Introduction

Recent progress in IP based optical backbone networks as well as broadband access environment such as digital subscriber line (DSL), CATV internet and fiber to the home (FTTH) allowed multimedia contents to be delivered to a wide scope of users. Wireless devices are adding the support for the standard Internet communication protocols to provide a rich application environment, which enables delivery of information and interactive services to digital mobile phones, pagers, personal digital assistants (PDAs) and other wireless devices.

Over such heterogeneous network and terminal device environment, contents need to be delivered seamlessly and in a manner to meet user's requirement and preference. A number of current content delivery services, however, are managed by provider's-side logic without knowing user-side requirements and/or usage environment. If the network knows user's profile, status and context, it will be possible to deliver contents in more convenient and suitable way for the user. Metadata plays an important role for above objectives.

This paper proposes integrated framework architecture of content delivery based on metadata. It provides a policy based content delivery control by using content metadata and user metadata. Two aspects; "Metadata driven QoS control" and "Metadata driven One Source Multi-Use", are discussed.

## 2. Metadata driven QOS control

Figure 1 shows metadata driven QoS control architecture. The basic idea of this architecture is to estab-

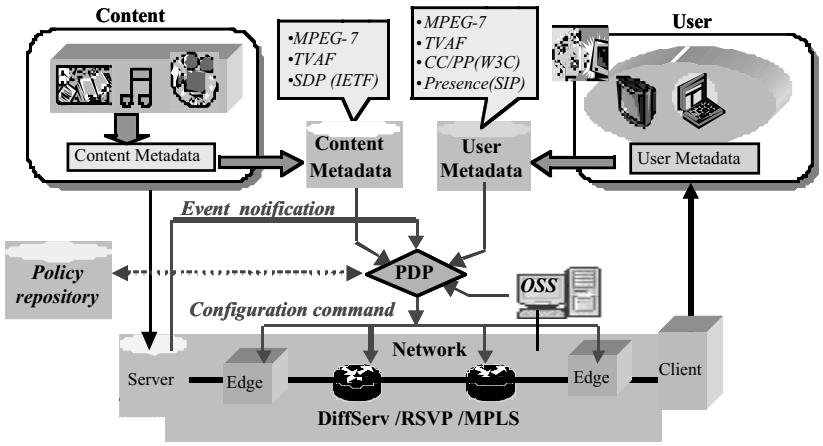



**Figure 1. Metadata driven Qos control**

lish feedback mechanism to reflect user side requirement in content delivery services. QoS level segregations such as quality assured content delivery or express delivery will be achieved, by harmonizing both content requirement and user requirement in network resource allocations through means of metadata.

Content metadata are defined in MPEG-7 [1] and TV Anytime Forum (TVAF) [2]. TVAF discriminates "content description metadata" that include content title, synopsis and genre, and "instance description metadata" that describes location, usage rules and delivery parameters. As for user metadata, user profile as defined in CC/PP (Composite Capabilities/ Preference Profiles)[3] of W3C includes terminal capabilities and user preferences. User metadata as defined in MPEG-7 includes user preference and usage history. Presence information as defined in IETF-IMPP (Instant Messaging and Presence Protocol) working group [4] includes user and/or terminal availability about participating in communications over networks.

In Figure 1, when a user requests a particular content to a server, event notification is sent to Policy Decision Point (PDP). The PDP refers to content metadata and user metadata to know about their attributes. The status of available network resource to deliver this particular content is obtained from Operation Support Systems (OSS) as needed. Then the PDP imports relevant policies that are stored in the policy repository to make a decision and sends configuration commands to relevant Policy Enforcement Points (PEPs).

## 3. Metadata driven One-source Multi-use

Recent rapid progress in wireless technologies is bringing ubiquitous service into reality. Internet access from mobile phone and/or personal data assistance (PDA) now allow computer and communication devices to continue communications even when mobile.

In Figure 2, a user who is viewing a MPEG-2 video by a personal computer at his/her office (or home) goes out and wants to continue viewing the same content by a PDA or a mobile phone. When the user switches the terminal device from PC to PDA (or mobile phone), the terminal device capabilities and access network conditions changes, thus arises the need of content adaptation to meet the new usage environment as well as user preference. Usage environment and user preference are provided by metadata and stored in user metadata database.

As examples of content adaptation, real-time transcoding, and source selection are envisaged. In real-time transcoding, the MPEG-2 video format is transcoded to MPEG-4 video format in a real-time basis to adapt to lower bit-rate of portable devices. MPEG-7 "MediaTranscodingHints DS" metadata can be used for this purpose. In source selection, the original video file is encoded in several video formats beforehand, so that an adaptation server can select a suitable source at content request depending on the terminal capability. MPEG-7 "Variation DS" metadata can be used for this purpose. These adaptation metadata are stored in the policy repository.

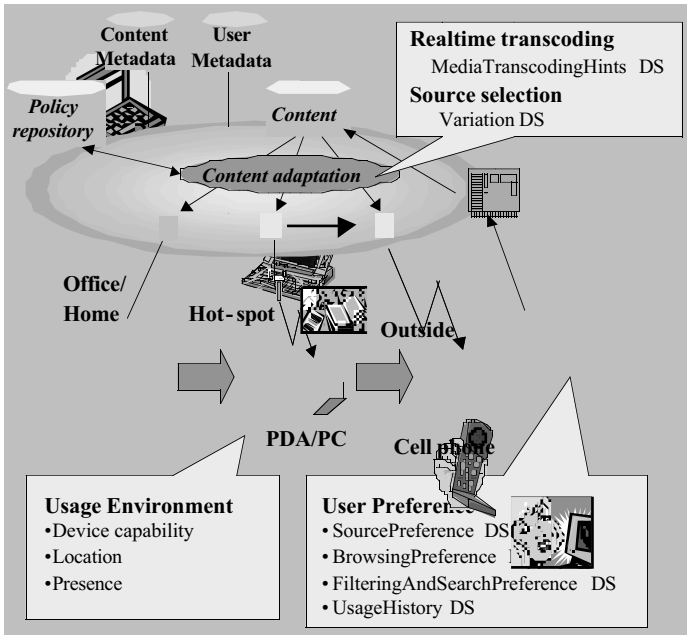



**Figure 2. Seamless roaming with content adaptation**

## 4. Conclusion

By using proposed framework architecture, content delivery service providers can establish particular policies of how to control content delivery by harmonizing various requirements described by metadata to achieve user-oriented services.

## References

[1] "Overview of the MPEG-7 Standard ", Dec. 2001.
[2] http://www.tv-anytime.org/
[3] W3C, " CC/PP", W3C Note 27, July 1999.
[4] "IMPP", IETF RFC 2779, Feb. 2000.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 229-234*


# Visual Representation and Contextualization of Search Results – List and Matrix Browser

Christoph Kunz, Veit Botsch
Fraunhofer IAO
Nobelstrasse 12
D-70569 Stuttgart
Tel +49-711-970-2334
{christoph.kunz, veit.botsch}@iao.fhg.de

## Abstract

*We present a new approach for the representation of search results in a graphical user interface that allows navigating and exploring these results. An interactive matrix display is used for showing the hyperlinks of a site search or other search queries in different hierarchical category systems. The results of a site search are not only shown as a list, but also classified in an ontology-based category system. So the user has the possibility to explore and navigate within the results of his query. The system offers a flexible way to refine the query by drilling down in the hierarchical structured categories. The user can explore the results in one category with the so called List Browser or in two categories at the same time with the so called Matrix Browser (Ziegler et al. 2002). A familiar and well known interactive tree widget is used for the presentation of the categories and located hyperlinks, so the handling of the system is very intuitive.*
**Keywords:** *Search engine, meta data, ontology, matrix browser, list browser, topic map, information visualization, classification*

## Introduction

Networked information structures are becoming increasingly important for exploring and navigating complex information spaces, such as Internet sites, knowledge repositories or engineering data. Information networks are also becoming important in the context of the Semantic Web (Berners-Lee et al. 2001) as metadata or ontologies for information indexing. Complex ontological information can, for instance, be expressed in formalisms such as Topic Maps (Biezunski et al. 1999) or DAML+OIL (Hendler 2001). Visualizing and exploring such network structures, however, still constitutes a major problem for user interface design, in terms of minimizing visual search, supporting user's understanding and providing efficient interaction for exploring the network.

The exponentially growing amount of information available for example on the internet, in an intranet or a file system increases the interest in the task of retrieving information of interest. Search engines usually return more than 1 500 results per query and the results are displayed in a plain list with only few meta information. In general, people have two ways to find the data they are looking for: they can search by entering keywords to retrieve documents that contain these keywords, or they can browse through a hierarchy of subjects until the area of interest has been reached. The two tasks of searching and browsing are separated in most of the search engines. The information located in the hierarchy of subjects is not used to classify and to display the search results.

The approach presented in this paper combines the two ways of searching and exploring information spaces in a new graphical user interface for search engines. Information units of the underlying information space are linked with the metadata layer as occurrences of the ontology topics (see Figure 1). Every information unit can be linked with different topics in different categories of the metadata structure. The ontology itself must be analyzed in a manner that structures with hierarchical or transitive properties can be automatically recognized and extracted (Ziegler et al. 2002), so the user gets the possibility to choose hierarchical parts of the whole ontology for representation of the search results. Choosing a specific domain by selecting a given hierarchical category system allows the user to refine his query and to get a structured result list. The list and matrix browser are used as front ends for the representation and navigation of search results, which are classified in an ontology-based, hierarchical category system. The results of a keyword search are prestructured by the system using the ontology-based metadata. So the user can navigate and explore the result set

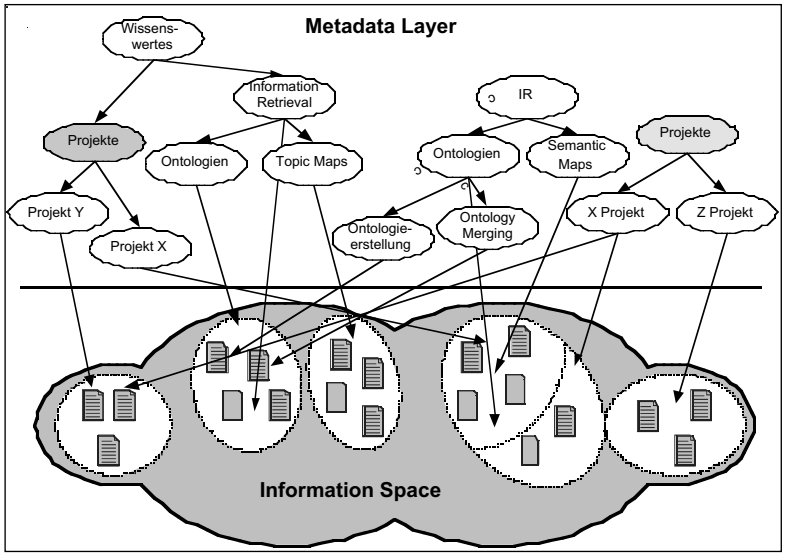



**Figure 1. The Metadata Layer and the Information Space**

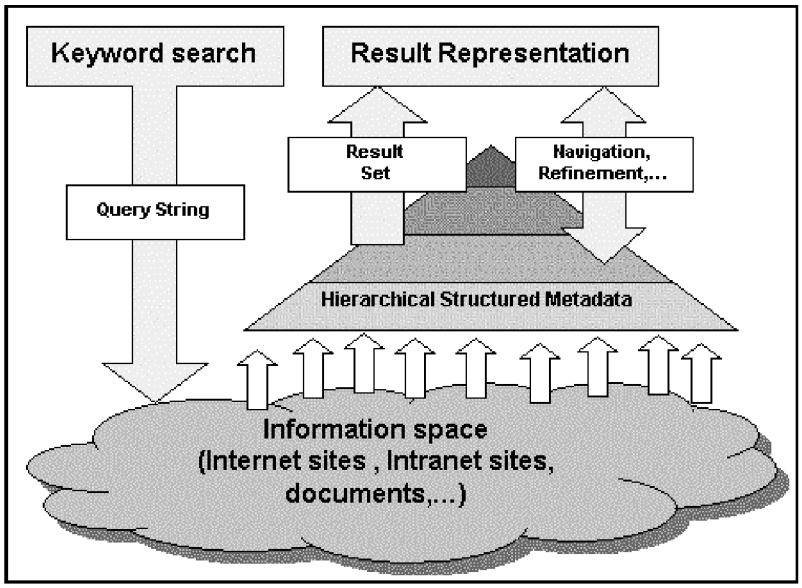



**Figure 2. The search process**

using the new graphical interfaces, without entering new queries (see Figure 2). The user can look at the returned hyperlinks with different, interactive "glasses", the chosen category system represents the glass.

The results of a keyword search are either displayed in one interactive category tree, the list browser or in an adjacency-like system with two interactive trees on the axis of a matrix, the matrix browser. In both cases the system provides a good review about the results and it offers the possibility to refine the query in a flexible manner. The first overview generated from the graphical user interface shows how many hits are found in the hierarchical ordered categories. If only one hierarchy is displayed, the number of hits is equivalent to the size of the displayed bar, if the user selected two hierarchies to structure the search results, the size of the squares inside the matrix encodes the number of hits belonging to the two categories. The user can explore the search result by expanding and collapsing the interactive, "window explorer" trees containing the hierarchical categories and the hyperlinks of the result set. If the user choose another hierarchical part of the category to visualize the search result, no re-querying is needed, because the result set is the same, only the representation has to be changed.

## Related Work

Hierarchical ordered category systems are used in several search engines in the web (e.g. www.yahoo.com, www.google.de, www.dmoz.org, www.altavista.com) to allow the user not only to search the internet by keywords, but also present structured hyperlinks which can be explored by users. A new kind of search engine which presents the results in a graphical manner and which allows query expansion using a metadata system can be found at www.kartoo.com (see Figure 3).

The USU KnowledgeMiner is a modular software product for the structure of topics and for rendering access to information in heterogeneous IT environments uniform. The meta data extracted from existing data sources are semantically linked based on the topic map standards ISO 13250 and enable access to information from one central point. The structure thus established is displayed graphically (see Figure 4). The user quickly obtains an overview of topics and is led to the desired information via the navigation and retrieval components

A search engine which use a personalized fuzzy ontology to refine queries is proposed by Widyantoro (Widyantoro 2001). Gauch et al. (Gauch et al. 1999) use a personalized ontology for a re-ranking and filtering of query results. The search system incorporates users' interest into the search process to improve the results by creating user profiles. The

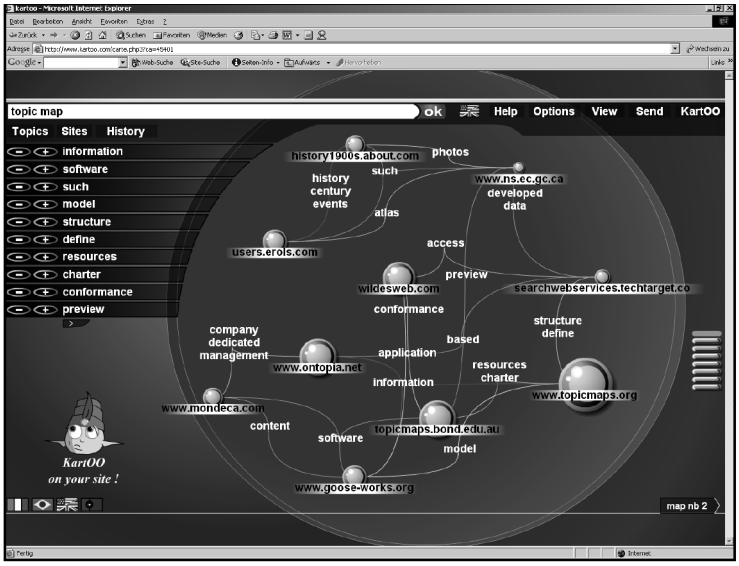



**Figure 3. Search engine www.kartoo.com**

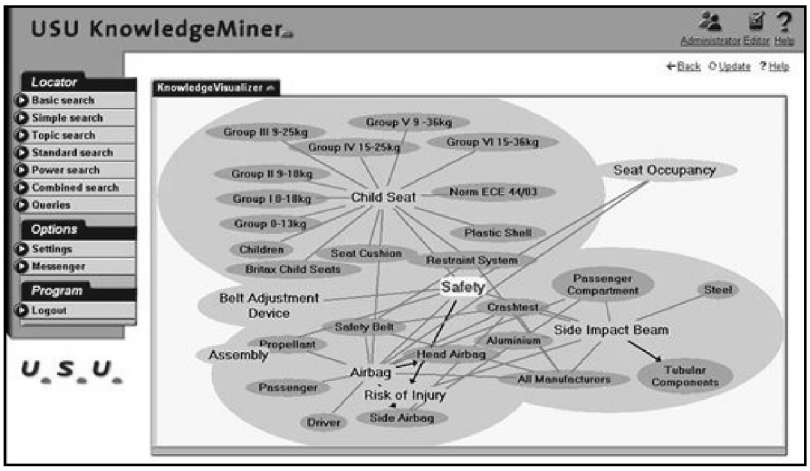



**Figure 4. The USU KnowledgeMiner**

user profiles are structured as a concept hierarchy of 4,400 nodes and are populated by 'watching over a user's shoulder' while he is surfing.

## System Overview

The idea of the List and Matrix Browser visualization is based on different strategies, like usage of known metaphors, high interactivity and user-controlled diminution of the amount of information by filtering mechanisms. The user has different options to reduce the shown information and also to increase the interactivity of the representation. One possibility is to narrow or to widen the search query by using the well known keyword search. Another possibility is to select specific category hierarchies, which are used to refine the query and to classify the result set.

Hierarchies placed along one (List Browser) or two axes (Matrix Browser) can be explored directly by expanding and collapsing their interactive trees. This kind of exploration is familiar and effective and works both on the horizontal and the vertical axis. By expanding and collapsing the familiar interactive tree of the hierarchies, the user can increase or reduce the displayed amount of information and refine his query.

## Requirements

The system needs an information space with a metadata network which is linked with the instances of the information space. Topics of the metadata structure contain occurrences (hyperlinks) of the information units. A site search engine with a category system is a possible domain for the new graphical user interface. Another potential usecase is a document management system which offers a structured hierarchical folder system and metadata belonging to specific documents. The results of searches in such systems with more than one hierarchical category system can be displayed either by the List or by the Matrix Browser. The information resource itself should be referred to in more than one category, other than in search engines like yahoo or google, where websites are only in one specific category.

## List Browser

If the information structure only contains one hierarchy of categories or if the user chooses only one hierarchy of categories, the results of a keyword search are displayed in an interactive representation of the categories, the List Browser (seeFigure 5). The categories, all subcategories and the hyperlinks to the sites of the result set are listed in a "Windows Explorer"-like tree widget, which allows the user to expand and collapse the categories and subcategories by clicking on the interactive symbols in front of the bar and the category name. Not all categories of the metadata representation are listed in the List Browser, empty categories, where no hits of the search are located, are not visible in the representa-

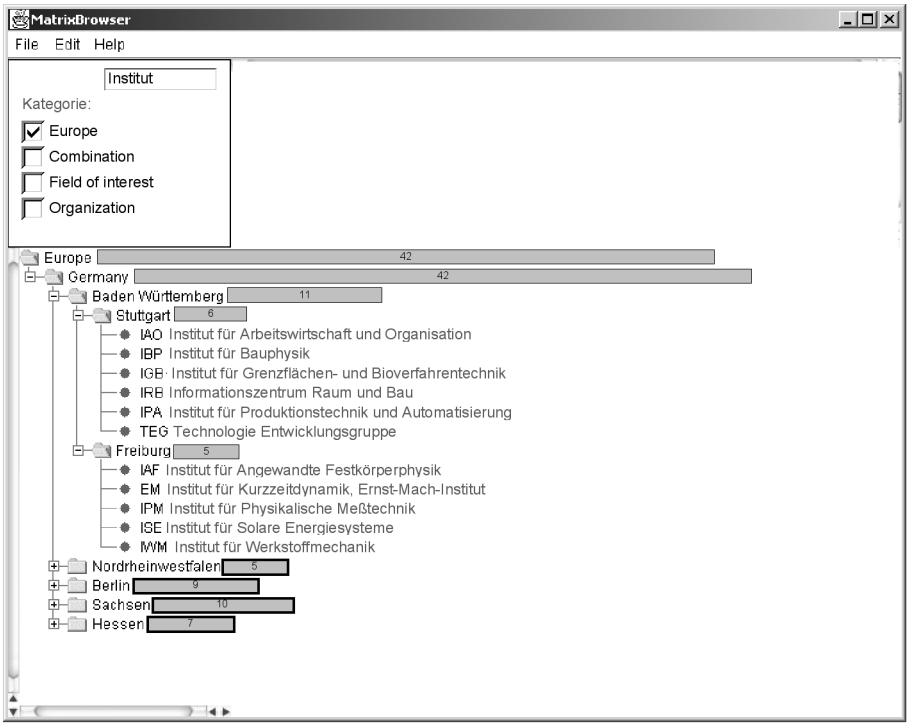



**Figure 5. The List Browser**

tion of the search result. The size of the bar represents the consolidated amount of search hits in the category and all subcategories. So the user gets an overview in which categories the search results are located and he can explore the results in a flexible manner.

## Matrix Browser

The Matrix Browser (see Figure 6) allows the user to display the result list in two different categories. Choosing two categories also refines the query itself, because only links in the chosen categories are shown. The chosen categories are displayed as interactive, "Windows Explorer"-like trees on the two axis of an adjacency matrix. The user has the possibility to navigate within these trees and to explore the structured search result without re-querying the information space. The interactive matrix gives a good review about the search result and the reference of the found sites to the metadata category system. Different kinds of interactive symbols inside the matrix visualize on the one hand how many sites are found in two categories (the size of the circles) and on the other hand the site reference itself (the squares). Like in the List Browser the hyperlinks to the located sites are listed in the interactive category trees. If the hyperlink is listed in both trees, the horizontal and the vertical, this hyperlinks are connected visually with a square if both hyperlinks are in the focus. If one of the hyperlinks is in a collapsed state the square contains a plus and works interactively, so the user can also click on the square to expand the appropriate category. If a site is located in both category systems, but both categories are in a collapsed state, then a circle is shown in the matrix. The size of the circle represents how many located links are in the hierarchy under the referred nodes. Some hyperlinks are only listed in one hierarchy, they don't have a link to the other category system. In Figure 6 the listed institute CLS has no link to the category "Europe", because the institute is located in America.

## Further Developments

Hyperlinks between sites of the search results could also be shown in the Matrix Browser, using another symbol for this kind of link. This additional information shows the level of networking between the located sites. We also think about including the consolidated values of located sites for the categories, already shown in the List Browser, in the Matrix Browser. An important factor for the backend system is the automatic generation of hierarchical structured metadata, which is linked with the information units.

First user evaluations of the new graphical user interface have to take place to improve further developments of the visual front end for search engines. The task of searching and browsing with the List and Matrix Browser has to be compared with other keyword-based search engines and category systems.

## Conclusions

This paper described a new graphical user interface for ontology-based search engines, that allow the user to navigate and explore the results in a familiar

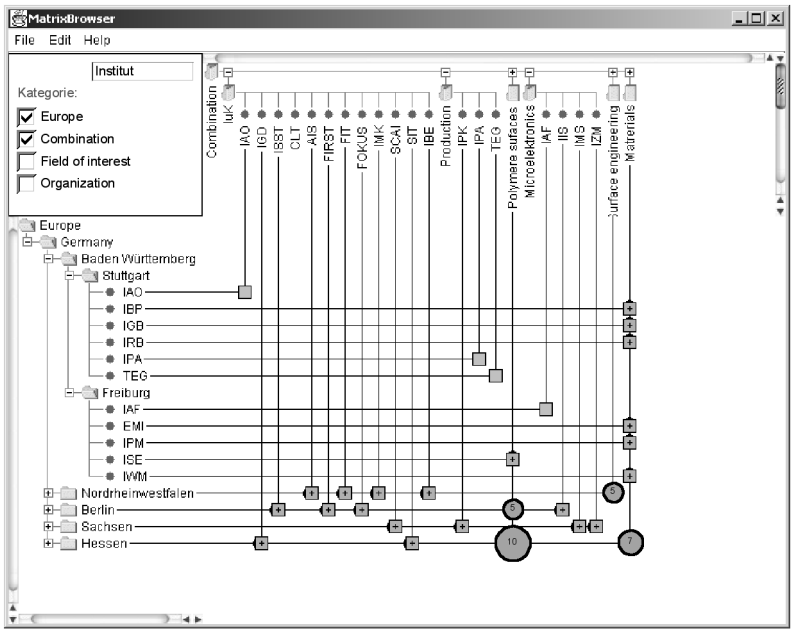



**Figure 6. The Matrix Browser**

and intuitive manner. Hierarchical category systems are displayed in interactive tree widgets, so the user can increase or reduce the displayed amount of information and refine his query without input of any text data. The two tasks of searching and browsing are combined in one graphical interface. Using the Matrix Browser provides an opportunity to visualize more details of the metadata structure together with the located sites. The result set of a keyword search is shown in a part of the metadata structure, so the user can chose different "glasses" (parts of the metadata structure) to look at the results.

## References

Berners-Lee, T., Hendler, J., Lassila, O. 2001. The Semantic Web, Scientifc American, Vol. 284, p 34-43.

Biezunski, M., Bryan, M., Newcomb, S. R. 1999. ISO/IEC 13250 Topic Maps: Information Technology — Document Description and Markup Language. http://www.ornl.gov/sgml/sc34/document/0058.htm.

Gauch, Susan; Pretschner, Alexander 1999. Ontology Based Personalized Search. In: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, Chicago, 1999, pp. 391-398.

Hendler, J. 2001. DAML: The DARPA Agent Markup Language Homepage. http://www.daml.org.

Widyantot, D.H., Yen, J. 2001. Using Fuzzy Ontology for Query Refinement in a Personalized Abstract Search Engine. In: Proceedings of Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Vancouver, Canada, 2001.

Ziegler, Jürgen; Kunz, Christoph; Botsch, Veit 2002. Matrix Browser - Visualizing and Exploring Large Networked Information Spaces, In: Proceedings of the International Conference on Computer Human Interaction SIGCHI 2002, ACM Press, Minneapolis, 2002.

Ziegler, Jürgen; Kunz, Christoph; Botsch, Veit; Schneeberger, Josef 2002. Visualizing and Exploring Large Networked Information Spaces with Matrix Browser. In: Proceedings of 6th International Conference on Information Visualization, London, England, 2002.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 235-236*


# Development and Application of Dublin Core Metadata Standards in Marathi

Shubhada Nagarkar
Sr. Technical Assistant
Bioinformatics Center, University of Pune, Pune – 411 007
shubha@bioinfo.ernet.in

Harsha Parekh
Librarian and Prof. of Library Science
SNDT Women's University, Mumbai - 400 020
harsha_parekh@vsnl.com

## Abstract

*To enable Dublin Core metadata to become a global standard for locating information on the Web, it is essential that such metadata be provided in different languages and in a variety of scripts. This would enable search and retrieval of Web documents in all languages. Considering the potentials of DC metadata in resource discovery on the Internet, the current paper reports an attempt to translate Dublin Core metadata elements into Marathi language (the local language of Maharashtra State in India), render them in Devnagari script and apply them to a significant section of Marathi literature – viz. the writings of one of the "saint-poets" of the Middle Ages.*

*Our efforts are to create a web based database and to assign Dublin Core international metadata standards rendered in Marathi for cataloguing the literature of one of the prominent "saint-poets" - Chokha Mela - available in print and electronic format as well as on Internet. This is conceived as part of a larger project of organizing all the literature of the "saint-poets".*

*We have chosen a group of "saint-poets" in Maharashtra from the 13th century (e.g. Jnandev 1275-96) to the 17th century (e.g. Tukaram 1608-90) who helped establish the 'Bhakti' (devotional) school of Hinduism in Western India and assisted in its spread from southern to northern India. Even today, the lives of the saint-poets and their literature continue to inspire a large section of the Marathi speaking population and provide them with emotional solace. As such, their writings constitute an important segment of Marathi literature and researchers from all over the world are engaged in studying it. The original writings in Sanskrit (including the early handwritten manuscripts) and commentaries in Marathi on them are scattered at different places in India and are now beginning to make an appearance on the Web.*

## Objectives of the study

- To translate the DCMES into Marathi (version 1.1).
- To test the translated elements with cultural heritage literature – the literature of the medieval poet saints.
- To find out limitations and add new qualifiers if needed.
- To send translated DC into international registries.

## Project description

The project of translation of metadata into Marathi and its application for saint literature (database available at http://bioinfo.ernet.in/~shubha/dc/main.htm) is divided into following phases:

**1. Translation:** Each of the DC element (with definitions and comments) was translated into Marathi. For this translation work we referred to various dictionaries and subject headings and consulted linguists, grammarians and literature experts. Table 1 shows the basic 15 elements and their translation into Marathi.

**2. Rendering them into Devnagari script:** Once the translation of the DC metadata elements was completed, the work of actual implementation began. Initially we have used Shiv01.ttf Marathi font, which is available free on the net and is phonetic as well. The work done using Shiv01.ttf font, is based on ASCII character and could not be converted directly into UTF-8 which is widely accepted Unicode standard. However, we are experimenting with the tools available at C-DAC, Pune, (www.cdacindia.com) which help to convert ASCII based fonts into ISCII based ones and then into UTF-8.

**Table 1. DC elements into Marathi (Version 1.1)**

| | | | |
|---|---|---|---|
| Title | SaIYa-k | Type | p`kar |
| Creator | inama-ata | Format | rcanaa |
| Subject | ivaYaya | Identifier | AaoLK |
| Description | vaNa-na | Source | ]gama |
| Publisher | p`kaSak | Language | BaaYaa |
| Contributor | sahBaagaI / sahayyak | Relation | naato / saMbaMQa |
| Date | idnaaMk / tarI#a | Coverage | vyaaPtI |
| | | Rights | h@k |

**3. Converting it into UTF-8 character coding:** We also found that Microsoft has developed a Marathi Unicode font - Mangal which is available with Windows 2000 and Windows XP. We faced problems in displaying the Unicode font. Efforts are underway to use Microsoft's Web Embedding Font Technology (WEFT) to convert the Unicode font into - Embeded Open Type (EOT). Further we are working to use Mangal.eot with special java script that enables users to see the Devanagari script on any platform viz. Windows 95, 98, 2000 and NT with any version of IE 4.0 + and Netscape communicator version 4.3+.

**4. Metadata registry:** The Marathi DC elements are being sent to the DC registry at OCLC, USA (http://wip.dublincore.org:8080/dcregistry/index.htm) and to the registry maintained at University of Library and Information Science (ULIS), Open Metadata Registry, Tsukuba (http://avalon.ulis.ac.jp/registry/).

## Discussions

Our primary attempt here was to translate the DC elements into Marathi. DC translation into Marathi was fairly simple and we found appropriate terms in Marathi. A blind reverse translation indicated that in almost all cases the Marathi term for the DC element was correctly translated back to the English term as well as properly understood. In some cases (e.g. identifier and relation) the understanding and the reverse translation indicated that it was necessary to provide the context by the detail qualifiers. Thus at the end of the process we have come up with DCMES Marathi translation (Unicode based) version 1.1(http://bioinfo.ernet.in/~shubha/element.txt). We are continuing to test these elements with a wider range of materials – including hand-written palm leaf manuscripts, music, films, etc. As the database grows to include this material on Chokha Mela and expands to cover other saint-poets, we will be able to test, improve and enhance the translation of DC metadata. Updated versions of this translation in Unicode fonts will be made available on the said web site and with the registries.

## Acknowledgments

Authors are thankful to the Bioinformatics Center, University of Pune for the assistance to carry out this work. We express our sincere thanks to Principal P.D. Puranik, Editor and Mr R.P. Kulkarni, Co-Editor of Dnyshewar Trimasik and of special volumes on Saint Chokha Mela, for their valuable help. Without the help provided by Mr R.P. Kulkarni we would have not completed the translation part. We are thankful to Mr M.J. Nagpurkar, Shri Mudran Mandir (printer) and president Dr M.G. Shahade of Brahman Madyavarti Mandal (publisher) of Saint Chokha Mela volumes. He helped us in converting all these articles into pdf files. We highly appreciate the help provided by Priya Informatics, Pune (Dealers of C-DAC products) for their help in conversion of fonts from ASCII to ISCII and to Unicode. We also express our thanks to Prof. Shegio Sugimoto and Mitsuharu Nagamori from the University of Library and Information Science (ULIS), Tsukuba, Japan, Prof. Shalini Urs, Mysore University for their consultations.

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 237-241*


# Why is Accessibility Metadata Proving Difficult?

Liddy Nevile
Motile Research
liddy@motile.net

## Abstract

*Accessibility metadata is simply metadata that describes the accessibility of resources and services, usually those on, or available through, the web. Awareness of widespread web content inaccessibility led to work being done to develop guidelines for authors and others to make sure that content would be more accessible to those with special access needs, especially those with disabilities who were being disenfranchised by their lack of access to the web. Currently, work is being done to find ways of signalling the degree of accessibility of resources, and ways of matching resources to searches and people. In addition, accessibility metadata could be used to repair some inaccessibility problems on the fly. This paper describes some of the work being done and the problems that have contributed to make the progress comparatively slow.*
**Keywords:** *Accessibility, metadata, Dublin Core, DC-accessibility, Evaluation and Report Language, EARL, people with disabilities, guidelines, W3C, IMS.*

## 1. Introduction

Accessibility metadata is, put simply, metadata that describes the accessibility of resources and services, usually those on or available through, the web.

Web content accessibility became a topic in the mid nineties. It was realised that much of the content of the new 'web' was not accessible to people who did not use standard web GUI browsers, the same technology that was making the web attractive and available to naïve computer users. Many people complained, at that time, that they could not download the 'much too big' files, or that the colours were not consistent, but a whole swag[1] of people suddenly found that the very technology that had enabled them to rejoin society was suddenly alienating them. In particular, blind people, people with motor coordination problems, in fact, many people including those who could not use a mouse on a computer screen for one reason or another, were suddenly not able to use their computers as their life-style-support machines. Additionally, people who depended, for one reason or another, on screen readers were often being read content that was unrecognisably jumbled, according to the GUI layout specifications of the author.

The World Wide Web Consortium (W3C, [1]) responded by establishing a Web Accessibility Initiative (WAI, [2]) program to work on what was making web content inaccessible. Since that time, W3C WAI have developed extensive guidelines as to how to make content accessible to all devices, so that those using non-standard combinations of hardware and software could access web content, or content available through the web, if their devices were standards compliant. This work is undertaken under the banner of the Web Accessibility Initiative, is open to all, and is international, partly due to its special funding structures. The immediate aim was to avoid content that would be totally inaccessible to some, before working on making all content more generally accessible.

The W3C WAI works on how to make offending content accessible, often by repairing it. They concentrate on what accessibility would be but also on the authoring and user access tools. The WAI Authoring Tools Accessibility Working Group [3] has emphasised how to make authoring tools, for instance, productive of accessible content, even when the author is not aware of what is necessary. Such emphases were chosen to get the greatest benefit to the greatest number as quickly as possible, in the realisation that authoring was becoming more complex and more people would soon be using authoring tools.

Repairing an inaccessible page; identifying inaccessible content; techniques for making accessible content, are all under control, if not yet completely documented. What is required now, is work on metadata to perform a number of roles. Of course, discovery is a primary goal. Finding a resource or service is an on-going problem on the web, and all the usual difficulties operate when people with special needs use

[1] Australian expression meaning what is rolled up and carried around as 'home' by tramps. It usually contains everything but the kitchen sink.

the web. Everyone has a need for information that suits their purposes at the time they seek it. This may occur in special circumstances, such as when people working underground in protective clothing, perhaps because they are working in a mine, need to access information (possibly how to deal with a leak), without using their hands, and so without keyboards. These users would possibly need to be able to use their voice-controlling software to use the command-key navigation of a web page. They will need to know if the page is properly constructed so that such navigation is possible. If it is not well-constructed, they may need to know:

- how it is constructed so they can determine if they will be able to get to the information in some compromised way, or
- if they can expect some transformation application to make it accessible or finally,
- if there is no hope of access for them.

The challenge is to find a suitable way of expressing and disseminating accessibility metadata and to make it available as soon as possible.

In this paper, we consider how the W3C Web Content Accessibility Guidelines (WCAG, [4]) and other guidelines, in a sense derived from the WCAG, have helped towards the problem of identifying how to produce metadata about resources and service accessibility, and what problems remain. Primarily, the author asserts that the rush to solutions[2], without time to establish a clear set of requirements, has left organisations interested in this metadata with the very difficult task of trying to fit requirements to solutions – known in the software industry generally as a serious nightmare situation! In many cases, the actual requirements will be known only at the time, and this makes it additionally difficult. Additional resources and style sheets, for instance, may also need to be retrieved to accompany the original resource.

## 2. Accessibility

### 2.1. Accessibility Information

Assessment of accessibility usually requires an assessor to identify what types of content are contained within a particular resource, and therefore which of the total array of guidelines apply, to what standard and thus, if the resource as a whole is compliant. At best, assessors then make an assertion about the accessibility or otherwise of a resource. (This is just another example of a situation in which it is important for the consumer of that assessment or evaluation to know who made it).

When single resources are to be evaluated, a report on their compliance is usually produced. As the number of resources increases, and the frequency of their evaluation, and the increase or otherwise of accessibility of collections becomes of concern, metadata management becomes an issue. Not only will people want metadata to discover the accessible resources, or to transform them appropriately, in addition document management agencies will want records and reports automatically generated about the accessibility of the resources, possibly integrated into their document management systems.

People can have special needs because they are temporarily in a situation, for instance one that makes their hands unable to work the keyboard, or the noise level so high they cannot hear a screen reader, or otherwise. In order to help people with special needs, it is necessary to identify their needs and requirements.

The most useful way to do this is to develop a normative set of requirements and then have users or agents select from that. This approach has been used: the banking industry, for example, has worked to accommodate selected people with special needs at their Automatic Teller Machines. The people fit particular profiles and are issued with smart cards that adapt displays for this purpose. Determining a comprehensive normative list is a non-trivial exercise, however.

Once such a list has been determined, the W3C WAI guidelines, for instance, can be matched to the requirements and relevant sub-sets of accessibility determined. This again is a major task, given that there are hundreds of checkpoints involved in compliance with the W3C WAI guidelines to achieve a general accessibility rating. The good news is that many of these checkpoints can now be tested automatically, at least to the level of partial compliance, or possibly more usefully, complete non-compliance. Some criteria, such as the provision of a text alternative to an image, is failed if there is nothing but not compliant unless what is there is intelligent.

### 2.2. Accessibility Solutions

The main accessibility solutions take a number of forms but three are notable:

- Guidelines as to what should and should not be published, or more precisely how content should be published, e.g., an image should not be published unless it is accompanied by an ALT (text alternative) tag, and, if the image conveys significant information that is required for comprehension of other content, it should be accompanied in addition by a long description, and that should be done in one of two ways, and so on.
  These guidelines, in the case of W3C WAI especially, have been developed after significant consultation and account for a range of access devices that may be in use, catering simultaneously for people with different devices and, incidentally, with a range of disabilities. In the case of W3C, these

[2] An activity in which the author has been engaged for a number of years.

guidelines are not model or agent specific, and are written to be robust under evolving conditions of the web and associated tools.
- Checklists [5] have been developed to allow assessors to work with some sort of consistency. These checklists provide more or less detail but aim to clarify what might be considered compliance, again differing in nature according to whether this information is to be understandable to a naïve assessor or only useful to an expert.
- Technique documents [6] aim to show that it is possible, and hopefully practical, for developers to author resources that will comply with the guidelines. They can be thought of, at one extreme, as proof-of-concept documentation, as there may be more ways of achieving the goal, but at the other extreme, these are tools for accessible content creation, The range and availability of such collections of techniques is extensive.

Finally, there are many situations in which organisations are developing what might be called accessibility policies. Making resources fully accessible can be burdensome, and may not always be appropriate. Organisations around the world are working on what is feasible in their context, what they will set as their local standard. Such policies often work by selecting criteria from the W3C WAI list. The guidelines do not go so far as to provide a specific, actionable set of requirements that can be immediately included in development specification documents. But more problematic is that neither do they provide for absolute testing of compliance: compliance is subjective, and for many of the guidelines, can only be assessed by an expert. Ultimately, of course, accessibility is totally subjective, and depends upon a user at the time. It should be remembered that accessibility as described in guidelines does not guarantee useability; that too has to be assessed subjectively by humans.

### 2.3. Accessibility Requirements

Imagine the needs of a deaf person.

A deaf person does not have hearing problems with content that does not have sound associated with it but in many cases, the spoken language used by the surrounding community is a second language for the deaf person. Such a person might be used to sign language, and so be a non-native speaker, and therefore reader, of text that is included in a resource. In some cases, text at all is not helpful to the deaf person, and they may require the text to be *translated* into sign language for them. The requirements for accommodating a deaf person are not simple, as shown, but depend upon the strategies adopted by the deaf person to operate with their disability.

Deciding whether a resource is suitable for a deaf person, if done by a third person, is a form of censorship. In many cases, this is not appreciated by people with disabilities: they would prefer to know what it is with which they might have difficulty, and then decide how much effort to make and what compromises are acceptable to them in the particular circumstances.

Even where it is necessary according to business or functional specifications for a resource to be classified as suitable for deaf people, as might need to happen if an organisation's accessibility policy is to ensure its resources are accessible to deaf people, it will not be a straight-forward matter. Accessibility technical requirements specify, *in technical terms*, what resources need to do and usually the developers have to determine the most appropriate way to achieve these aims. This is the ideal situation and likely when professional developers are at work, using the most rigorous techniques. It is not what happens in most cases in practice.

And, as shown above, there are so many variables and dependencies that it may be better for the deaf person to be enabled to say what they want, by choosing from what is available, than to be subjected to some automatic feeds.

Such considerations often lead to a call for description of people's needs and sometimes, on to descriptions of people's disabilities. Again, this is not considered an appropriate approach in all circumstances, especially as it can easily degenerate into breaches of privacy, and error.

## 3. Resource Matching

### 3.1. Matching Requirements to Solutions

Metadata that describes the accessibility or otherwise of a resource in particular circumstances is likely to include a lot of information and not be the sort of thing a person will just read, in the way they might have read a paper library catalogue record. It is not likely that something as simple as 'accessible' or 'not accessible' will be meaningful.

In fact, in order to promote accessibility, and reward those who tried, W3C introduced a scheme of levels of accessibility, A, AA and AAA. Unfortunately this proved over-simplistic. Compliance with all but one small detail in a set of criteria meant failure to that level, even if there was also compliance with almost all the other criteria necessary for compliance with a higher level. This proved discouraging and not helpful. A new system has been proposed for the future. Other organisations promoting accessibility have tried rewarding authors with icons [7] but these have been abused by ill-intentioned and misused by ignorant[3] people to the point where the icons lack credibility. Anyway, they lack precision and are not very useful to users with specific needs.

[3] A typical example is offered where people try to evaluate sites using the Bobby test and do not read the fine print and assume their site is accessible when Bobby just says it would be accessible if …

This leaves a situation where applications are required to handle all the accessibility metadata. In such a case, inter-operability is also important and as part of that, the semantics and syntax used for the metadata. In addition, as the quantity of metadata increases, due to continual testing and other document management activities, metadata management will become important.

Fortunately, the work of W3C on syntax in the context of Resource Description Framework (RDF, [8]), used already by many for metadata, has led to a useful development that will help solve some of these problems. Evaluation and Report Language (EARL, [9]) is a derivative of RDF and has the form:

A asserts that X was compliant with Y on Z date

Such a statement leaves it up to the user, or agent, to decide on the trustworthiness of such a statement but makes it possible for a computer to interpret the statement and, if necessary, act on it. In such a case, the computer may be using discovery software but equally, transformation or other applications or even searching for complementary resources, such as captions files.

### 3.2. Matching Resources and Services to Users

Bringing together users and resources, as has been mentioned, is a bit like dealing with censorship. How best to do it is probably more of a political issue than a technical issue, but nonetheless difficult for that. Technically, the choices relate to the difference between server-side and client-side solutions. Servers can restrict what is made available to users:

- during the discovery process;
- at delivery time, or even;
- from within a composite resource that has alternative content that is intended to be varied according to a request of the user agent seeking it.

In some situations, users will not want to receive content they cannot use because they will have telecommunication constraints that make redundancy expensive in terms of time, money, and in other ways. In other situations, possibly the same person will want to receive complete resources, like everyone else, in order to maintain privacy about their needs.

User agents, acting on the client-side, can modify requests before making them to a server, or filter what is received and present only portions to the user.

Client-side technology that immediately pre-dated RDF, designed for such a purpose, was the Platform for Internet Content Selection (PICS[4], [10]). To use PICS, a user selects from a range of options made available on a form by their user agent, and that information is converted into a numerical representation to be used by the user agent to control user presentations. PICS was extended to include semantic and structured values and values for repeated variables and transformed into RDF.

PICS remains, however, as a technology that might be useful in this context. The IMS Global Project (IMS, [11]), a consortium of those interested in developing tools for education, especially in situations where a student's path through the use of resources and assessment points is to be monitored, have adopted the approach of using a 'Learner Information Profile'. It is the IMS' intention to add some accessibility requirements into this profile. This will, somehow, be matched with resource accessibility metadata that will be attached to the resource.

### 3.3. Developing the Vocabularies

Finally, just as with any other classification system, different organisations will have local purposes and requirements and will want their metadata to support those needs. This means that there is a distinct likelihood of the different agencies wanting to use different sets of elements to make up their vocabularies for their metadata. One of the factors that makes it 'hard' to work in the field of accessibility metadata is that there are not already accepted and tested vocabularies. There are not even keywords in common usage. Accessibility metadata is a new idea.

W3C WAI, having the most comprehensive set of criteria for accessibility evaluation, is working towards a numerical schema for identifying compliance with individual criteria. Such a schema could provide a useful resource for others who then could merely pick and choose among the criteria for their local purposes. It would also promote inter-operability of accessibility metadata.

Computer-Useable Metadata

Another quality of accessibility metadata that is not unique but is challenging, is that this metadata is designed to be used by both computers and people. So far, there are not applications waiting to retrieve inaccessible resources to turn them, on-the-fly, into accessible resources. The enabling technology is already developed for this, however. In determining the metadata formats to be used, it is obviously important to have this potential in mind, but as it is not clear what will be required, or more practically, used, it is difficult to decide what to do about it.

## 4. Metadata Associations

### 4.1. Dublin Core and Accessibility Metadata

Finally, determining how the Dublin Core Metadata Element Set (DCMES, [12]) can be used to provide accessibility metadata is not yet settled.

[4] Later, PICS was associated with censorship, most particularly because it could be used at the proxy level, and so has not been a 'popular' technology.

One of the charter aims of the DC-Accessibility Interest Group [13] is:

- to determine the relationship between accessibility (evaluation and repair) descriptions and DC descriptions - and report on an appropriate way to represent accessibility information in DC.

Specific questions to be answered are:

1) Is accessibility a resource discovery issue?
2) What is the relationship between accessibility (W3C's EARL) descriptions and DC descriptions?
3) Is it sensible to embed one in the other?
4) Could one provide, as part of DC RELATION, information about the relationship between equivalent resources?
5) Should EARL schemata be recommended?

There is little debate about the value of accessibility metadata for discovery although there is some about how the requirements should be matched with the resource's qualities. Accessibility descriptions will not be about the content of the resource, but rather one of its qualities. This suggests that it is not DC-description type information, and so would not be well located there. Nor does this information make a lot of sense as DC-format information. It is not about the format, but how it is used.

The question of whether DC-relation is a good place for information about equivalent resources for use by those who cannot access the primary resource is not so clear. If a resource is well-developed, from the accessibility perspective, it will have information or alternative resources correctly embedded or associated with it, so the use of DC-relation will be the same in this context as others, to identify something that may be an interesting alternative resource. Where the original resource is not well-formed, and the user is dependent upon the replacement of one element or all of it by another, not originally associated with the primary resource, this information should be in the metadata relating to the primary resource. Otherwise, where what is required is the use of a different style sheet, for instance, this will need to be retrieved and the original resources related to it instead of whatever was originally proposed. In either case, the accessibility of the resource will be changed by the effect of this metadata, and the post-hoc provision of the equivalent element. This information should be closely associated with the description of the accessibility of the primary resources, as it affects it. For this reason, it makes more sense to have accessibility and alternative and/or equivalent resource elements co-located. So it is probably better to have a separate accessibility element with all this information in it.

EARL descriptions will be RDF expressions but deciphering metadata in RDF that is well-formed with clear parsing rules should not present a problem in a DC environment. One argument for embedding EARL statements, or other accessibility descriptions, in DC metadata is the potential for wide-adoption of accessibility metadata if it is associated with the incredibly wide-spread DCMES.

If DCMES is seen as a suitable affiliation, or association for accessibility metadata, the question remains as to how this will work out in practice. It can not be mandated that people must produce and use accessibility, or any other, metadata. Nor can it be predicted with accuracy whether they will. It is hoped, however, that if the Dublin Core Metadata Initiative can take the lead on this, with the collaboration of other interested bodies, that whatever is chosen as the DC-accessibility solution, will be useful to and popular among others concerned with this issue.

### 4.2. The way Forward ...

What seems possible is for those working on accessibility metadata to work together. 2002 has been a year of integration and currently W3C, IMS and DC-Accessibility have brought together their activities to the individual benefit of all groups, as well as for the general problem.

EARL is reaching maturity, and expected to be promoted as a W3C recommendation by the end of 2002. IMS Metadata specifications, particularly the Learner Information Profile, designed to be a metadata management tool for tracking student progress and needs, is reaching maturity and also expected to be ready for recommendation by the end of 2002.

## 5. Conclusion

Although all metadata standards are 'hard' to develop, particularly as their global uptake depends upon local utility, as well as a number of factors to do with inter-organisational and international cooperation, accessibility metadata standards, in particular, are 'hard' to achieve.

## References

[1] http://www.w3.org/
[2] http://www.w3.org/wai/
[3] http://www.w3.org/wai/au/
[4] http://www.w3.org/WAI/Resources/#gl
[5] see eg http://www.w3.org/WAI/Resources/#ch
[6] see eg http://www.w3.org/WAI/Resources/#te
[7] see eg http://bobby.cast.org/
[8] http://www.w3.org/RDF/
[9] http://www.w3.org/2001/03/earl/
[10] http://www.w3.org/PICS/
[11] http://www.imsproject.org/
[12] http://dublincore.org/
[13] http://dublincore.org/groups/access/

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 243-244*


# Subject Access Metadata on the French Web

Ewa Nieszkowska
École nationale des sciences de l'information et des bibliothèques
Lyon, France
niesz@enssib.fr

## Summary

The article presents four French projects on subject metadata use: a medical portal (Caducee.net[1]), a subject gateway (Les Signets[2]), a catalogue of patents (INPI, Institut National de la Propriété Industrielle[3]) and a full-text database of the daily newspaper "Libération". The last project is not a public Web application yet but it presents the most innovative approach to subject metadata usage discussed in the article.

These projects, both completed and in progress, as a common characteristic share the use of controlled documentary languages. By this means, they try to increase the efficiency of information retrieval for the remote user.

The article tries to determinate the "remote user" characteristics: he or she is defined as a person searching for information (often for professional purposes) and who often needs exhaustive information in the situation of chronic time shortage. The most popular search engines cannot satisfy such users, who need a more organised Web, and more efficient search. In fact, they might also need a librarian, although they do not know it yet!

However, when they sit alone facing their computer screens, they do not receive assistance from information retrieval specialists (as, for example, librarians). In this situation, it is the role of a resource provider to help remote users in their documentary search and to make this search more user-friendly.

How do the studied projects approach this problem? For Caducee.net and Les Signets, it is done by means of a fairly classical use of indexing languages. In the case of Caducee.net, it is achieved by the use of a standard familiar for the medical public called MeSH[4] (F-MeSH in French). Les Signets face the problem by the planned use of RAMEAU[5], French indexing standard, which has very large number of "used-for references", i. e. non-descriptors that can guide the user to descriptors themselves and, in this way, to relevant resources.

INPI case is more interesting and unusual. Since the indexing language is an alphanumeric one (complex class symbols incomprehensible for a remote user), a linguistic engine is employed to enable search in natural language. Afterwards an index of keywords is generated from existing verbal descriptions of class marks themselves.

All the above-mentioned projects show the importance of natural language tools for remote users. And the fourth project's study seems to indicate that the use of controlled languages in full-text environment can be beneficial for controlled languages themselves: it's the case of the daily "Libération" thesaurus.

This thesaurus appears to prove that full text documentary environment may also be used to create and/or maintain indexing vocabularies and thesauri. The descriptors of the thesaurus are associated (via KALIMA software) with "lexical units" from the full-text articles database. Of all the thesaurus' modules, two seem particularly interesting, as they make an inventive use of an association between "lexical units" and the thesaurus' descriptors. They are called the "Automatic Learning Module" (ALM) and the "Automatic Indexing Module" (AIM):

ALM works by extraction of texts selected for learning, linguistic analysis of their contents, comparison of the contents with their indexing descriptors, finally by saving the results of the comparison in "indexing prototypes". Every time the ALM is used, it generates a new "indexing prototypes". At the same time, the thesaurus' administrator is asked to validate or reject new associations created between the words and expressions coming from the text and the thesaurus' descriptors. The recommendations for validation or rejection of these lexical units are based on their frequency in the given text.

AIM's function is to draw up a list of relevant candidate descriptors of new documents that have been put in the database. This process works by extraction of all the documents' fields (i.e. text title and the body), linguistic analysis of the contents, comparison

with the thesaurus and then final comparison with indexing prototypes of AIM. As the output, the librarian receives predetermined number of the closest descriptors whose relevance has been assessed by the AIM. Afterwards, the librarian's work is to validate, or reject the candidate descriptors and to add those, which have not been generated by the system.

It is important to understand, and underline, that the creation and maintenance of indexing languages is one of the library activities that incur the highest cost. It stems from the fact that for the time being it has been impossible even to part-automate it. The example of the "Libération" thesaurus seems to be opening up a different perspective - not for documents on the Web, but rather for documentary languages …

… Full text contribution is to reduce tedious human workload in the maintenance of indexing standards. It is for the machine to take care of scanning texts and then to compare them with the existing descriptors. This way, the process of assembly, selection and choice of indexing vocabulary, as well as its maintenance, is considerably accelerated. The librarian is made to take final decisions, but whatever can be automated, will be. To put it simply: do not ask what metadata can do for the Web; ask what the Web can do for metadata.

---

[1] http://www.caducee.net
[2] http://www.bnf.fr/pages/liens/index.htm
[3] http://www.inpi.fr
[4] Medical Subject Headings
[5] Répertoire d'autorités-matière encyclopédique alphabétique unifié, http://rameau.bnf.fr

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 245-250*


# What's the Use of DC.Type? Semantic and functional aspects of the role of DC.Type within a moving image metadata generation tool

Simon Pockley
Australian Centre for the Moving Image
simonp@acmi.net.au

## Abstract

*Type has come to the fore as one of the primary organizing elements in the design of input forms suitable for the generation of high quality moving image metadata. A lack of semantic precision in both the definition and in the conceptual complexity of DC.Type's encoding scheme has prompted a re-evaluation of its usefulness as an element to be populated for interchange and discovery. In order to introduce precision to this element, a distinction is made between subject-based descriptors (genres), object based descriptors (forms), and manifestations or format-based descriptors (formats). A DCT2 vocabulary is proposed for DC.Type as a point of discussion for facilitating the deployment of domain specific encoding schemes and for filling gaps in the current list of terms.*
**Keywords:** *DC.Type, type, genre, form, Semantic web, Metadata interoperability, Cultural heritage metadata, Knowledge management.*

DC.Type is one of the purest metadata elements, in so far as it is a term directly associated with our inherent need to order, categorize, classify and group similar resources. Yet, the benefits of semantic precision have been elusive for this element. Within the diverse reaches of the Dublin Core community, it has long struggled to find uncontested territory of its own. The many and various applications of the term, 'type,' have worked against it having a simple set of values. In its current form, sitting uncomfortably in the nether regions of the borders between DC.Format, DC.Relation and DC.Subject, 'type' could be an attribute of any element and therefore suffers from having to do too much.

DCMES 1.1 defines Resource Type as, the nature or genre of the content of the resource (being described). As it stands, the values of DCMI Type Vocabulary make up a coarse-grained but conceptually complex and semantically troubled list. This list compresses a single level of aggregation, several subjective descriptions or 'genres', and a mixture of high and low level physical format designators or 'forms'. The current approved list of terms consists of:

- Collection
- Dataset
- Event
- Image
- Interactive Resource
- Service
- Software
- Sound
- Text
- Physical Object (proposed)

The DC.Type Working Group's archives[1] trace the origin of this encoding scheme and provide fascinating insights into the various ways that reservation and important assumption can drop out when time is short and a schedule of deliverables takes precedence. There is no intention here, to be critical of the valuable work of this group. This group attempted to reconcile the natural, non-exclusive, non-hierarchical structures of usage with the forcing of unnatural, exclusive resource categories into a hierarchical classification scheme. The current encoding scheme is defended as a minimal, high-level list where low-level types can be extracted from domain, or application-specific, fine-grained terms. However, even at a high level, the useful application of such a complex mixture of terms is proving to be a semantically daunting task.[2]

In common usage, the term, 'type' is often used interchangeably with 'genre' and sometimes 'form' or 'nature'. It is used as a loose way of signaling a descriptive shift to a different level of aggregation. For example in a classical music web site that explains different musical forms, we can see a distinction between three levels of aggregation (my parentheses). The cantata is described as 'an important genre (level 1) of vocal chamber music':

> *Secular cantatas in German and Italian were composed by Keiser, Telemann, Bach and others, but this type (level 2) was never cultivated to the*

*extent it was in Italy. In France and England the secular cantata was essentially an 18th-century genre, (level 3) emulating the Italian type (level 2). (Boynick 2001)*[3]

In the arts, the various patterns of critical interests that have drawn on Aristotelian poetics as a way of aggregating works into types have achieved their status not because they fit together into any preconceived system or taxonomy, but simply because they recur constantly and independently. Literary theory is littered with the ruins of genre definitions that have convinced no one save their author. Communities of interest generally apply such terms as a pragmatic convenience where the act of categorization has occurred within a tradition of continuous redefinition.

The practical challenges of categorizing some of the more complex forms of new media that have appeared in the electronic, the digital, and the networked domains, are being addressed by communities that share an interest in managing the moving image in a range of analogue and digital formats. If, in the digital domain, they lag behind their text-based colleagues, it may be due to the twin challenges of complex technical dependencies along with massive file sizes.

Amongst members of the moving image community, as well as the movie consuming public, 'genre' has been, at once, the most useful method of grouping film and video, as well as the most deconstructed and conceptually unsound method of classification. This has arisen from an attempt to establish the credibility of media studies through an assertion of seriousness and separation from the less weighty entertainment values of Hollywood cinema. It has also come from a need to provide a grouping mechanism for the continuous production of top lists both as an aid to discovery and, by inference, to establish criteria for assessment or interpretation. After a century of film production, the conventions of 'genre' are also being used as stylistic shorthand as well as being an inherent component of the production of meaning. Like all literary forms, moving images constantly refer to themselves and to other cross-media generic manifestations.

In order to find firm ground on which to base a rationale for populating DC.Type, as an element with a consistent encoding scheme, it is useful to reach back into the origins of European thought and apply the triple distinction made by Aristotle between description by subjective response, by words, and by mimicry/imitation. This becomes a useful mechanism for distinguishing subject-based descriptors (genre) from object based descriptors (form), and manifestations or format-based descriptors (format).

This is not new. It is, in essence, the approach taken by Brian Taves (Chair) Judi Hoffman and Karen Lund in the *Library of Congress Moving Image Genre–Form Guide*. The guide uses MARC-based cataloging conventions to build up tri-part (genre-form-format) descriptions of moving image works. Notions of 'genre', and 'form' are described as follows:

*Genres are recognizable primarily by content, and to a lesser degree by style. Genres contain conventions of narrational strategy and organizational structure, using similar themes, motifs, settings, situations, and characterizations ...*

*... Forms are defined as the basic categories indicating a moving image work's original exhibition and release parameters (such as length and medium), and which are separate from its actual content, not necessarily implying a particular narrative construction. Form terms include Feature, Short, Serial, Animation, and Television, and can be associated as needed with any genre, in a manner similar to free-floating subdivisions ... While the form indicates the work's original appearance, a third field, format, such as film, video, or videodisc, indicates the actual physical characteristic of any particular copy. For instance, a videodisc of THE SOUND OF MUSIC would have the genre-form-format heading "Musical—Feature—Videodisc". (Taves 1998)*[4]

Responsive, non-linear forms might usefully be added to the Moving Image Genre–Form Guide. These would include such terms as: web-site, gameplay, generative, installation, interactive, simulation, surveillance, and ambient works. These are all forms that are (or can be) dynamic and open in nature. Library and archival communities have tended to avoid collecting examples of such works because they are difficult to capture except by 'snapshot'.

A useful test for 'form' is that form is an objective description with a precise but repeatable value. For example, a work described as a 'short' may also be an 'animation.' Whereas values for genre are imprecise, subjective terms with many shades of meaning that might be adapted to critical purpose such as documentary, film noir and crime.

Most of the values for the encoding scheme of DC.Type are, by this definition, high-level forms. Low-level precision will come with the ability to apply domain specific values for forms consistent with the notion of objective definition.

A semantic distinction between form and genre offers a level of precision that is missing in the approach taken in the *Metadata Object Description Schema* (MODS).[5] In this schema, 'type' functions as an element level attribute. For example, 'genreType' has form values: motion picture, newspaper, periodical, picture, video recording, web site etc.; 'typeOfResourceType' has form values: text, cartographic, notated music, sound recording, still image, moving image, three dimensional object, software, multimedia, mixed material etc.; and 'physicalDescriptionType has form as a subset along with 'internetMediaType' and 'extent' and is given the enumerated values of: Braille, electronic, microfiche,

microfilm (similar to DC.Format).

While genre terms might have limitations as subject heading values, communities who use and augment pragmatic applications of LCSH for discovery purposes would not find much difficulty in accommodating their own genre schemes into DC.Subject. Genre lists are by no means exclusive to moving images. Since 1991, Medical Subject Headings (MeSH)5 has listed 'Publication Types' to describe 'forms' of presentation. At its lower sub-type levels, genre terms are used to describe materials based on their cultural or literary forms such as addresses or sermons or their physical forms such as broadsides or posters.

At its higher 'form' level it is curious to note that the MeSH encoding scheme is considered to refine DC.Subject rather than DC.Type when they share terms.

## Populating DCType

Type is a grouping attribute that could be applied to almost any DC element. When it comes to discovery, the challenge is to be able to extract information from rich records in a way that can be expressed using DC elements without disrupting inter-application interoperability. For practical discovery purposes (assuming an XML/RDF syntax), 'type', as a conceptual notion or display, rather than as a compounded element, could be retrieved or populated from a rich metadata record by an aggregation of the values of attributes from different elements. Assuming either DC.Type or DC.Subject had the benefit of the refinements of a separation between genre and form:

- DC.Type (domain vocabulary) Form + DC.Subject (domain vocabulary) Genre + DC. Format (domain vocabulary) Medium

or, through a qualified version of DC.Type:

- DC.Type (domain vocabulary) Form + DC.Type (domain vocabulary) Genre + DC. Format (domain vocabulary) Medium

For example, moving image, feature, DVD, or text, lecture, pdf

When the current DC.Type vocabulary was first proposed, the Working Group for DC.Type recognized that greater precision would be achieved by using more specific descriptors, but rejected the concept of multipart expressions on the grounds of 'retrieval considerations'.

> *… We expect additional structure for values of DC.Type to emerge from forthcoming discussion, allowing greater granularity of resource types to be expressed within this overall framework. This is likely mainly to involve sub-typing, for example including terms to indicate such things as moving vs. still images, different types of text, etc. However, the structure and syntax of Qualified DC has not been resolved at this time. A refined structure for Type will be implemented according to the general recommendations for Qualified DC. (Cox 1998)*[6]

Somehow between the separation of the minimalist approach expressed in Simple Dublin Core and the unrealized refinements of Qualified Dublin Core this form of semantic precision was lost to DC.Type. At DC9 in Tokyo, the DCMI Type Working Group decided that it would not try to produce an 'official DC' subtype list, and that such lists would be created by domain-specific working groups or by applications.

Conformance with the approved values of DC.Type involves transforming and extracting terms from lower level schemes and including the values of other elements.

## What's wrong with DC.Type?

1. No encoding scheme registration process: A domain specific registration process is planned.

2. Image is too coarse: As a term, 'Image' needs some form of refinement. It currently covers any visual representation other than text such as photographs, paintings, prints, drawings, diagrams, maps, musical notation, animations and moving pictures, film, VR/3D environments and is a sacrifice to minimalism that compromises the usefulness of the term. The moving image is one of our major expressions of cultural heritage. At the very least still images and moving images should be separately defined.

3. Obscure terms: the term dataset (once data), as defined, stands out as belonging to the language of a technical community and might be replaced by 'template' as a term with more general currency.

4. Misplacement: 'Interactive Resources' may well have represented the zeitgeist of 1999 after the production of CD-ROM 'interactives' but in 2002 this term might be more usefully categorized as a subtype of a responsive or dynamic mode of encounter amongst terms such as:

- ambient works
- environments VR/3D
- game play
- generative
- installation
- interactive
- simulation
- surveillance
- web site

5. Element overlap: The term 'Collection' was proposed because of the need to identify a collection without describing its parts. Aggregations such as 'collection' are already expressed in DC.Relation.HasPart and one of the most important characteristics of

the Relation element is that it has an item level corollary Relation.IsPartOf. In practice, the term, 'Collection', by itself, is not nearly so useful because hierarchical trees or relationship models cannot be generated. Currently, the use of an implied default value to describe an item is obscure and an unnecessary complication to any encoding scheme. There is no question that the ability to describe a resource as a collection is needed. That this should be asked of DC.Type is worthy of challenge. The proposal to add 'Aggregation-level' to DC.Type by the DCMI-Government Working Group opens up the more complex issue of how to express levels of granularity.

From a moving image perspective, it is worthwhile noting that emerging standards such as MPEG 7 and MPEG 21 provide the syntax for describing sequence, shot, frame and even elements within the frame. With the aid of appropriate image recognition tools, these standards have the potential to turn all moving image items into collections.

From a discovery standpoint, the reason we aggregate works is to make it easier to get at the parts. In addition, usage of the term 'collection' is anchored in the Library or Museum community and confuses people who see themselves as building exhibitions, programs and packages rather than collections.

## What's the use of type?

One of the unique exhibition spaces of the Australian Centre for the Moving Image (ACMI) is its Screen Gallery. This space, converted from two underground railway platforms into the world's largest digital media gallery, will feature the most innovative of Australian and international screen-based art, including:

- responsive installations
- large-scale projections
- video and computer animations
- interactive works
- net art
- immersive environments.

The juxtaposition or montage of film, television and multimedia will encourage multiple interpretations of themes, and an understanding and appreciation of how the various media interrelate.

The primary goal in outputting metadata conforming to standards such as the Dublin Core is to be able to exchange records with others and to expose selected fragments of this metadata for global exchange. These records also provide a source of content for footnote screens in the screen gallery, back-of-house administration, printed catalogues, reports, displays, lists, things to do, audio tours, interactive experiences as well as control over the flow of information about valuable assets (including the metadata itself). A significant departure from the item/format centred model of our main legacy database was to base the notion of what constituted a 'chunk' of information on the David Bearman model[7]. In this model, works are expressed in many forms and/or performed at many times and may be produced in numerous manifestations. Each metadata record is based on the intellectual content of the work rather than on its particular form, manifestation and format. For example a video postcard work by Robert Cahen entitled Cartes Postales can be expressed in a linear form as a short and be manifested as VHS video in PAL or it could find expression as a non-linear multi-screen two-channel installation in MPEG2 at 6 mbs.

The changes that networked digital technologies have made in the way digital content can be produced and, by implication, discovered and consumed are most evident at the point of creation. They have already resulted in some significant changes in the management of audio-visual content, regardless of format.

**1. Shift from passive consumption to active use/production**

Cheaper digital moving image production tools (such as the iMac) combined with in-built encoding software are leading to increased screen literacy with an explosion of rich media content. We are also beginning to see tools with meta-logging software built in to the production and editing cycles (e.g. Sony Tele-File). It is important that such tools and content management systems are flexible enough to be able to be integrated with other systems. Yet the end-to-end approach of vendors forces a proprietary dependency anathema to collaborative or distributed activities and metadata exchange.

**2. Shift in managing multiple manifestations**

The re-purposing of rich media content goes beyond proprietary obstructions to cutting and pasting combinations of audio-visual fragments. Often multiple manifestations are required of a single work to suit different outputs and configurations. In the face of rapid developments in encoding software, it is important to attempt to store master files of uncompressed content from which different encodings can be made. Such content is unlikely to be exposed for public consumption. Depending on your point of view, rich media outputs are often manifested in unsuitable formats such as film, video, low-resolution codec, proprietary and even redundant formats.

While many cultural institutions are embarking on expensive digitisation projects for legacy content it would make sense to know who else holds the same resources and if they have already been digitised.

**3. Shift in identifying the borders of the work**

In a primitive way, the web has created a contextual universe around almost anything we can identify

with text. We now expect to investigate resources related to works that we may have previously viewed in isolation. The placing of borders around chunks of content has become a source of contention, about to be compounded by the wider deployment of RDF. The borders of a work have become as conceptual as the notions of what constitutes a collection.

Similarly, the reach of an Application Profile may soon define the borders of a business or a cultural or educational institution. In such spaces contextual resources are as inseparable from the notion of the work as the idea of it - in space and time.

These changes combine to create combinations of rich and domain specific metadata schema suitable for discovering complex digital resources. We presented a paper at DC9 in Tokyo outlining some of our experiences and practical difficulties encountered in the collaborative cataloguing of a wide range of digital artworks. Since then, the 'buy in' of curators and programmers has come through the development of different 'views' of our metadata generation engine or catalogue. Members of the D.C. Community who have had experience within cultural institutions will understand that exhibition oriented Curators and Programmers (key metadata creators) have quite different views of resources than Collection Registrars, Librarians and Conservators.

Our main cataloguing tool is a metadata engine that adapts itself to the perceptions and language of a range of users by providing them with different views of the record and its component outputs.

In an ideal world, the generation of high quality metadata begins at the point of creation. However, ACMI is a cultural institution that engages in creating or producing exhibitions and programs, commissioning works; and acquiring works by donation, purchase and internal production. This means that the process of metadata generation begins at the point of accession. For donated collections and failed production encodes, this can sometimes mean that the first metadata created is actually a record of deaccession. The point of accession or ingestion or disposal can vary according to whether a work is entering or leaving our collections.

A view of the record, tailored to the inputs needed to complete it, is activated by the selection of an Accession type from an administrative schema.

- Exhibition (a collection created by Curators)
- Program (created by Programmers either collection or item level)
- Event (created by Programmers either collection or item level)
- Production (internal – either collection or item level)
- Commission (external – either collection or item level)
- Purchase (includes donation either collection or item level)
- Loan (either collection or item level)
- Disposal (either collection or item level)

ACMI's Application Profile uses an XML/RDF syntax to augment and populate Dublin Core elements from a range of fine grained elements and attributes relating to the cross referencing of:

- Descriptive metadata: textual and visual documentation e.g. clips, stills, artist's statements etc.
- Interpretive metadata: e.g. exhibitions, programs, rationales, curatorial statements, interpretive essays, reviews, genres etc.
- Expressive metadata: technical requirements e.g. equipment lists, instructions, layout plans etc.
- Physical/production metadata: format and display descriptors e.g. aspect ratio, resolution, signal, frame rate, audio encoding, bit rate etc.

Terms selected from an unapproved DC.Type encoding scheme are used as triggers for displaying the appropriate elements to be populated. In our case, the conditional use of the high level form, 'moving image' can determine the values of attributes needed for recording complex Format descriptors for over 90,000 titles.

The categories of information needed to manage a range of MPEG2 moving image manifestations are quite different from those required for the time and place of an event which has no format; or text; or the dimensions or location of a physical object.

## Reworking DC.Type into DCT2

This paper acknowledges that different domains have quite different terms and needs. However, as a way of addressing some of the more restrictive consequences of an hierarchical approach to arranging the values of an encoding scheme, usage and broad representation would suggest several small but pragmatic changes to DC.Type. The following alterations would greatly improve the useful application of DC.Type and the consistency of lower-level encoding schemes:

1. an adjustment to the DC.Type definition to replace the word 'genre' with 'form' where form is described as an objective description of the resource;
2. an adjustment to the DCT1 encoding scheme to include only terms that are forms;
3. splitting the term image into the two high level terms, 'still image' and 'moving image'. This may require the DC.Type encoding scheme DCT1 evolve to DCT2 where the moving image and the still image are recognised as distinct top level terms with the definitions:
   ***moving image:*** *Definition: Any image created in a film, video, or other media format that alters with time and that is able to be displayed or projected on a screen. For example, movies, animations, television, multimedia, games, emerging media, simulations.*

**Table 1. Table of proposed changes**

| Current DC.Type scheme | Proposed DC.Type scheme | Rationale |
|---|---|---|
| Collection | | Resolve to DC.Relation.HasPart |
| Dataset | Template | More common usage suggestion |
| Event | Event | |
| Image | Still Image | |
| | Moving Image | |
| Interactive R. | Responsive Resource | Suggested (possibly dynamic) |
| Service | Service | |
| Software | Software | |
| Sound | Sound | |
| Text | Text | |
| Notation | Suggested possibility | |
| Physical Object | Physical Object | |
| | Web site | Added term |

***still image:*** *Definition: the content is primarily symbolic visual representation other than text. For example - images and photographs of physical objects, paintings, prints, drawings, other images and graphics, diagrams, maps. Note that image may include both electronic and physical representations.*

1. adding the term 'web-site' as a top level term;
2. removing the aggregating term, 'collection' from the scheme and resolving it within DC.Relation as a term that need not have its parts described;
3. recognizing that 'interactive' is now a lower level term of a form that is responsive or dynamic;
4. expediting the registration process for domain specific encoding schemes.

[1] DC.Type Working Group archives 1999-2000. http://www.mailbase.ac.uk/lists/dc-type/archive.html

[2] Agnew, Grace. V-Access@Listserve.UTK.EDU message sent 2002-05-31 thread: difference between 'genre' and 'type'?

*... Somehow or other we never ended up discussing this recommendation fully in our user guide. We also didn't see a useful place to put genre because, on the one hand, while it can be subject, our scenarios of use genres really aren't subjects. DCMI seems to use genre as synonymous with format, and that wasn't appropriate. Finally, we settled by default on the catch-all data element, "description".*

[3] Boynick, Matt. The Classical Music Pages: Musical Forms – Cantata. Last Revision - 10 October 2001 http://w3.rz-berlin.mpg.de/cmp/g_cantata.html

[4] Taves, Brian et al. 1998 The Moving Image Genre-Form Guide. Library of Congress Motion Picture/Broadcasting/ Recorded Sound Division, February http://www.loc.gov/rr/ mopic/migintro.html

[5] Metadata Object Description Schema (MODS) http://www.loc.gov/standards/mods/

[6] Cox, Simon et al. Type Element Working Draft. 1998 http://www.dublincore.org/documents/1998/10/23/type-element/

[7] Bearman, David et al. 1999 A Common Model to Support Interoperable Metadata. http://www.dlib.org/dlib/january99/bearman/01bearman.html

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 251-255*


# Using Dublin Core for DISCOVER: a New Zealand visual art and music resource for schools

Karen Rollitt, Adrienne Kebbell, Douglas Campbell
National Library of New Zealand Te Puna Mätauranga o Aotearoa
P.O. Box 1467, Wellington, New Zealand
Karen.Rollitt@natlib.govt.nz

## Abstract

*Discover is a web resource supporting the visual arts and music curriculum in New Zealand schools. It contains 2500 multimedia items from the collections of the Alexander Turnbull Library, which holds the national cultural heritage collections, and 300 resources from other sources. The product uses a metadata scheme that combines simple (unqualified) DC and qualified DC, EAD and local extensions expressed in XML and uses the RDF framework proposed by DCMI for expressing qualified DC in RDF/XML. This metadata schema will continue to evolve to support interchange of the NLNZ's digital resources within the library, archival and education communities.*
**Keywords:** *Discover, interoperability, Dublin Core, XML, RDF, Curriculum, Schools, Arts, Music*

## Introduction

The Discover Project supports the music and visual arts curriculum in New Zealand. It began as a pilot project for the Digital Library Programme in 2000. The objective was to select and digitise items to support the Visual Arts and Music disciplines of the Arts curriculum. The National Library of New Zealand plans to digitise resources for other curriculum areas and present them on Discover. This paper covers the:

- context for the development of Discover
- the standards used
- the items being described
- the metadata schema for the items
- syntax for the metadata: XML / RDF
- the application
- an overview of Discover

## 1. Context for the development of Discover

Discover was created as the pilot site for the National Library of New Zealand's digital collection. The primary goal for this collection is to ensure interoperability and interconnection through the application of standards to enable:

- sustainability over time
- access for those with disabilities
- the ability to retain rights holdings and permissions
- create data that would be exchangeable across platforms
- an authenticated reproduction of original
- and create standards based metadata that would support administration, resource discovery and presentation activities.

## 2. The Standards used

To enable interoperability the National Library of New Zealand's Discover Project put into practice its Metadata Standards Framework [7], which was published in 2000. This Framework includes standards ratified by national and international standards organizations, such as NISO and ISO e.g. Dublin Core and ISO23950; those ratified by the World Wide Web Consortium (W3) e.g. XML and RDF; and other widely used de facto standards such as jpeg and mpeg formats.

The Discover Project used many well-known standards as well as some of the new and emerging standards. Best practice and recommendations were also used especially for mapping and syntax. The DCMES is primarily used at the resource discovery level.

## 3. The documents described

2,500 multimedia items were selected from the Alexander Turnbull Library, which holds New Zealand's documentary research and heritage collections. These collections contain original material such as photographs, drawings and prints, oral histories, manuscripts and archives, and printed material

including books, newspapers, maps, magazines, and ephemera relating to New Zealand and the Pacific. The documents selected for Discover include paintings, photographs, posters, video clips, music, essays and bibliographies, created by New Zealanders and many reflect our Maori heritage. The original items were manipulated for web presentation and in many cases included an extract or a portion from the original item, e.g. a portion of a sound clip. Most of the items were single files (e.g. .jpeg, .tiff, .ra, .mp3, and .mpeg) rather than parts of a collection.

## 4. The Metadata Schema for Discover

Much of the metadata for the items in Discover was sourced from two existing National Library catalogues including, TAPUHI, a proprietary based database containing unpublished items, and the MARC based National Library catalogue. The metadata was exported from these catalogues and mapped to DC [8] and qualified DC [2]. All items needed additional metadata.

Discover required that the metadata support:

- resource discovery: DC (unqualified)
- resource description: DC; qualified DC; EAD; local elements
- and preservation, technical and administrative needs.

The metadata scheme selected for resource discovery is primarily DC (unqualified). For resource description qualified DC, an EAD element and some local extensions were used. The DC and qualified DC data was implemented strictly conforming to the DCMI recommendations, the non-DC extensions were added in a new namespace "nlnzdl" ("http://www.natlib.govt.nz/dl#"). The RDF schema provides a full description of the National Library's application profile [1]. Extensions to DCMES including their use are summarised in Table 1.

The EAD[3] Element, Digital Archival Object Location < DAOLOC> was chosen as an alternative to dc:relation because of the additional information needed for each of the surrogate files in order to present it intelligently. A DAOLOG tag can specify the URL, its role, its behaviour and a title. For each record there is a thumbnail view, a preview view, the online reference version of the object and the original object.

Metadata for the long-term management of the digital object itself, which includes information such as the digitisation process used, the size of the file and modification to the original is currently stored in a separate system.

A multi-step process is used to convert the metadata - firstly to DC in XML, then to DC in RDF/XML, and then the local extensions are added to the RDF (see figure 1). This allows delivery of extracted meta-

**Table 1.**

| Element Name | Attributes | Use |
|---|---|---|
| ead:daoloc | ead:role; ead:behavior; ead:locator; ead:title; | Digital Archival Object Location - the location of a remote Digital Archive Object. |
| **Element Refinement** | **Name** | **Use** |
| dc subject | nlnzdl:category | For subject browsing. |
| dc:identifier | nlnzdl:pid | Persistent identifier – a permanent name for a resource. The NLNZ uses the Handle System for its PIDs – naming authority 1727. |
| | nlnzdl:object | Location of a digital object. Duplicate of the Persistent Identifier in a commonly accessible format e.g. http rather than hdl. This is an interim solution until Web browsers can natively present handles with multiple locations. |
| | nlnzdl:local | An identifier used locally by an application, which may not be unique globally. |
| **Encoding Schemes** | **Name** | **Use** |
| dc:subject | nlnzdl:NZCT | Curriculum topics list for Discover. |
| dc:type | nlnzdl:LCTGM2 | Library of Congress Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms Headings. |
| | nlnzdl:LCSHFormOf Composition | Library of Congress Form of Composition |
| dc:identifier | nlnzdl:ATLNo | A local number used as the Alexander Turnbull Library Archival Collections Reference Number. |
| | nlnzdl:CAC | National NLNZ Corporate Art Collection Reference Number Scheme. |

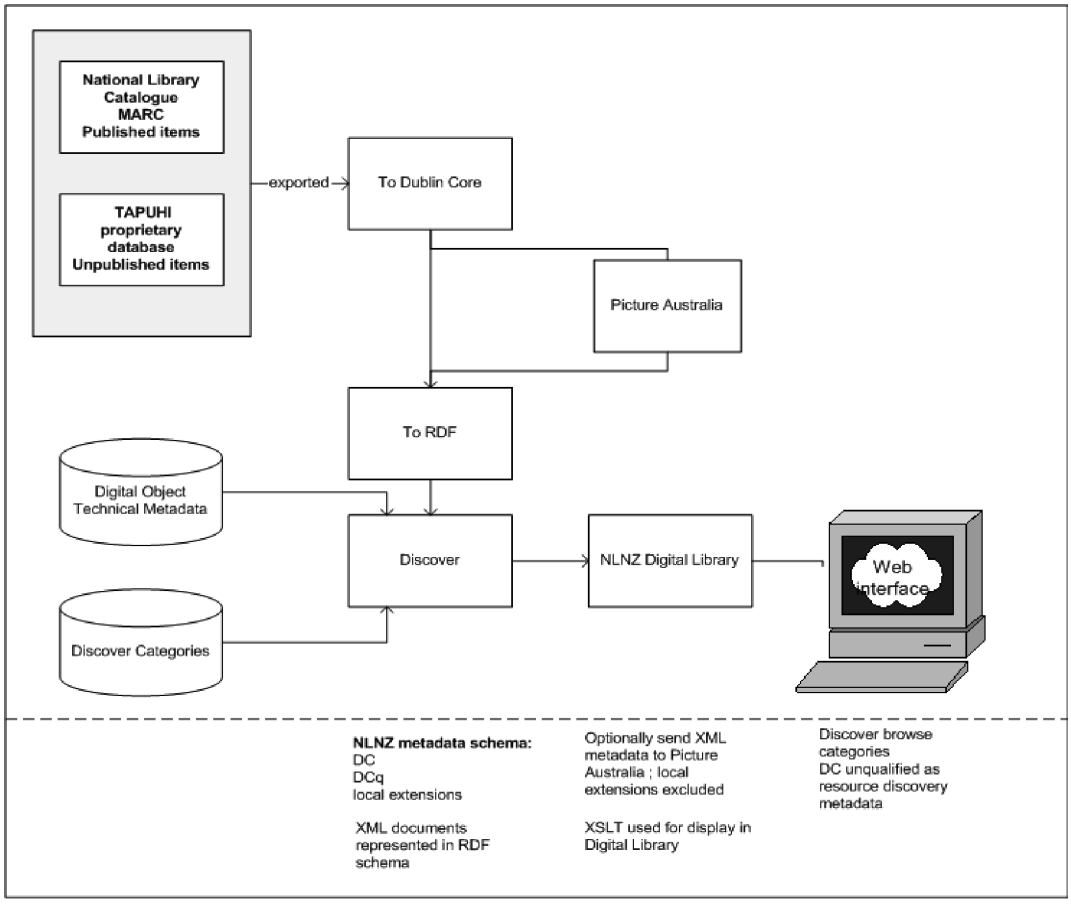



**Figure 1. Discover overview**


**Figure 2. Discover homepage**

data in different syntaxes depending on the destination's requirements. For example, the data sent to the NLA for inclusion in Picture Australia [10] is the result from the first conversion to DC in XML.

## 5. Syntax for the metadata: XML /RDF

The metadata for Discover is expressed in XML and the different schemas are combined using the Resource Description Framework (RDF). The NLNZ declaration [1] has been modelled on the DCMI proposal [5].

A DTD defining the Resource Description Framework XML for Discover metadata was also created because it is required by the NLNZ Digital Library application.

## 6. The NLNZ Digital Library and Discover

The NLNZ Digital Library application is capable of storing and both the import and export of data in eXtensible Markup Language (XML). It uses XSLT style sheets to interrogate the XML for display via the Web.

Discover is arranged into 13 topic areas to support the Visual Arts and Music Disciplines of the Arts/Nga Toi curriculum.

Retrieval is based almost entirely on DC (unqualified) although advantage is taken of some qualifiers to avoid confusion, for instance the kind of date being searched.

Figure 1 provides an overview of the processes used to generate the Discover metadata and is followed by an illustration of a Discover Web page presenting a stored digital object.

## References

[1] Campbell, Douglas. National Library of New Zealand Digital Library Dublin Core Enhancements RDF Schema. http://ead.natlib.govt.nz/meta/nlnzdlRDFschema.html

[2] Dublin Core Qualifiers. 2000. DCMI. http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/

[3] EAD Encoded Archival Description. . http://lcweb.loc.gov/ead/

[4] Kebbell, Adrienne. Digital standards. Paper presented at the National Digital Forum, held at the National Library of New Zealand, July 2001. http://www.natlib.govt.nz/en/whatsnew/forum.html

[5] Kokkelink, Stefan and Roland Schwanzl. Expressing Dublin Core in RDF/XML. 2002 http://dublincore.org/documents/2002/04/14/dcq-rdf-xml/

[6] Miller, Paul. 2002. A framework for access to the nations. http://www.natlib.govt.nz/files/forum/miller_files/frame.htm

[7] National Library of New Zealand. 2000. Metadata Standards Framework for National Library of New Zealand. Wellington, NLNZ, 29 p.

[8] National Information Standards Organization (U.S.). The Dublin Core Metadata Element Set : an American national standard / developed by the National Information Standards Organization. Bethesda, MD NISO Press
2001. ANSI/NISO Z39.85-2001. 6 p.

[9] National Library of New Zealand. 2002. New Zealand Register of Digitisation Initiatives. http://www.natlib.govt.nz/rodi/rodi.html

[10] Picture Australia. National Library of Australia. http://www.pictureaustralia.org/nolan.html

## Appendix. Sample Discover metatdata


```
<?xml version="1.0" encoding="UTF-8" ?>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcq="http://purl.org/dc/terms/"
xmlns:ead="http://www.natlib.govt.nz/dl#" xmlns:nlnzdl="http://www.natlib.govt.nz/dl#">
- <rdf:Description rdf:about="hdl:1727.11/00002195">
  <dc:title>Blue wattled crow (Kokako).</dc:title>
  <dc:subject>Kokako</dc:subject>
  <dc:subject>Birdsongs</dc:subject>
- <nlnzdl:category>
- <nlnzdl:NZCT>
  <rdf:value>A-M-04-04</rdf:value>
  <rdfs:label>Music and Bird Song Clips</rdfs:label>
  </nlnzdl:NZCT>
  </nlnzdl:category>
  <dc:description>Excerpt from track 2 of compact disc Bird calls of New Zealand. Paraparaumu, N.Z. : Viking Sevenseas NZ,
[1980]. Field recording by John Kendrick; narration by Robert Taylor; produced for the Department of Conservation and
Radio New Zealand.</dc:description>
  <dc:contributor>Kendrick, John L., recording engineer</dc:contributor>
  <dc:contributor>New Zealand Department of Conservation</dc:contributor>
  <dc:contributor>Radio New Zealand</dc:contributor>
  <dcq:issued>[1980]</dcq:issued>
- <dc:type>
- <dcq:DCMIType>
  <rdf:value>Sound</rdf:value>
  </dcq:DCMIType>
  </dc:type>
  <dc:format>Digital stereo sound recording, 59 seconds.</dc:format>
  <nlnzdl:pid rdf:resource="hdl:1727.11/00002195" />
  <nlnzdl:object rdf:resource="http://hdl.handle.net/1727.11/00002195" />
- <dc:language>
+ <dcq:ISO639-2>
  </dc:language>
  <dcq:hasFormat>Also available as an electronic resource.</dcq:hasFormat>
  <dcq:temporal>1980</dcq:temporal>
  <dc:rights>Item provided by the Alexander Turnbull Library, National Library of New Zealand, Te Puna Matauranga o
Aotearoa. Reproduction rights do not belong to the Alexander Turnbull Library. It must not be reproduced in any way without
the prior permission of the copyright owner and the Library.</dc:rights>
  <ead:daoloc ead:role="source" ead:href="http://digital.natlib.govt.nz/source/20020605/cd40track01_00002195_ds.wav"
ead:behavior="audio/x-wav" />
  <ead:daoloc ead:role="reference" ead:href="http://digital.natlib.govt.nz/20020604/cd40track01_00002195_df.mp3"
ead:title="Digital audio extract from Blue wattled crow (Kokako). (754KB)" ead:behavior="audio/mpeg" />
  <ead:daoloc ead:role="display" ead:href="http://digital.natlib.govt.nz/20020604/audioicon_pv.jpg" ead:behavior="image/jpeg"/>
  <ead:daoloc ead:role="thumbnail" ead:href="http://digital.natlib.govt.nz/20020604/audioicon_tn.jpg"
ead:behavior="image/jpeg" />
  </rdf:Description>
  </rdf:RDF>
```

*Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 257-259*


# A Proposal for a Flexible Validation Method for Exchange of Metadata between Heterogeneous Systems by Using the Concept of MicroSchema

Jens Vindvad
Riksbibliotektjenesten, Oslo Norway
Jens.Vindvad@rbt.no

Erlend Øverby
Conduct AS, Oslo Norway
erlend.overby@conduct.no

## 1. Introduction

A new method to solve the validation problem that arises when exchanging information between heterogeneous systems is proposed. The problem of validation is addressed by introducing the concepts of MicroSchema, used in a namespace environment.

To be able to share information between different systems, a well-defined protocol for information exchange must be in place. XML (Bray et al. 2000) has emerged as a new protocol for use in information systems for exchanging information between different systems.

Some of the challenges, when importing metadata from one system to another, are described in the experience learned by iLumina (McClelland et al. 2002) when importing IMS metadata. An issue reported was the need of validating against XML-model and error checking of imported metadata.

Normally two alternatives exist to describe and define the information structure or model in an XML document, the first is a DTD (ISO 8879) and the second is an XML-schema (Thompson et al 2001). Both these approaches currently have the disadvantages that in order to validate and check the structure of the information, description of the whole structure and all its possibilities and constraints must be in existence in one large and inflexible model, making it harder to establish an efficient validation of data exchange between different systems.

One reason almost everyone is using XML in only well-formed manner – is the flexibility in generating the information structures, if a new element is needed – it is just added and the information structure is still well-formed. Validation is often sacrificed. The disadvantage of only well-formed structures is that almost any element can be included, and there is no control of what the element names are or of their semantic meaning.

## 2. Conflict between rigid structures and the need for flexibility

When working with structured information, there is a conflict between flexibility, and the need for a rigid structure. If we try to look at the structure we normally find in a book, we will se that in many of our content models there is many similar structures. Normally parameter entities is used to manage that flexibility, but there is still a need to change the structure and to create new version of the DTD's. When using schemas to describe the structures, the notion of "global" element definitions can be used, but there is no function for describing content models in a flexible and reusable way. If wanting to change a content model by adding some new elements, is has to be done in a many different places in a schema, and only at once in the DTD.

One of the nice new features with XML over SGML is the introduction of the Well Formed document – which has the implications that there is no need to have a specified structure defined for the XML-document. This gives a great flexibility in processing the XML-documents and normally this is sufficient when there is full control of the information, and the processing of it. But if several people or systems producing information there is a need for greater control over the structure of the information that is produced.

## 3. MicroSchema

The challenge is to combine the flexibility in the well-formed document, with the control of the valid document. Using MicroSchema's this flexibility can be provided. The idea of a MicroSchema is that it should only describe a very small piece of information, and only such information as is relevant to the

specific description. Information that is not relevant to the specific context is described in another schema. MicroSchemas combines the flexibility of only well-formed documents with the need to specify and validate complex structures. To be able to express the relevance and the connection between MicroSchemas, a standard method of enhancing the schema specification in order to address the valid elements in the specific context is needed. Using namespaces, introducing the term "Allow-schema-namespaces", will do this.

Instead of specifying the whole structure in one or more schemas, only a small part of the structure in its own Schema (MicroSchema) is specified. Then the URI's is used to specify parts of the flexible Content Models. To some extent Parameter Entities can be looked upon as a URI reference from the MicroSchema. And the specification of Content Model or of the Generic Identifier (GI) is defined at the target URI. The URI will also work as the Namespace specification of the Semantic meaning of the GI's.

MicroSchema URI can bee addressed in two ways; one is as the Content Model specification, where one specific MicroSchema file is addressed in the URI.

```
<xs:element name="*" msc:gi="http://www.rbt.no/xmlns/
cerif/output/misc/chapter.msc"/>
```

*Example 1 Using the MicroSchema attribute GI*

In example 1 the xs:element will get the GI and Content Model of the element specified in the MircroSchema addressed at the URI http://www.rbt.no/xmlns/cerif/output/misc/chapter.msc. At the other hand only the Content Model could also be specified, using the MicroSchema specification for one element as shown in example 2.

```
<xs:element name="kapittel" msc:cm="http://www.rbt.no/
xmlns/cerif/output/misc/chapter.msc"/>
```

*Example 2 MicroSchema specification for one element*

In example 2 the element name "kapittel" will get the same Content Model as the MicroSchema specified at the given URI. Here this will replace the CHAPTER GI specified in the chapter.msc MicroSchema with the GI KAPITTEL given as the value of the name attribute.

The MicroSchemas and the corresponding documents are valid XML documents, and therefore can be processed as such. One of the primary ideas behind the MicroSchema is the XML-Well-formed processing, which does not require a set of rules against which to check the structure of the information. All XML MicroSchema documents are at least well-formed. The idea of a MicroSchema is to have the possibility of combining both well-form-ness and strict structures where the structure is expressed in a MicroSchema. Introducing the following three forms of MicroSchema processing rules does this: simplest form, simple MicroSchema check and complete MicroSchema validation.

## 4. CRIS as a test case

A lot of work has been done in the field of metadata exchange. Particularly initiatives like Dublin Core, Open Archive Initiative and work with Learning Object Metadata (LOM). To demonstrate and test the concept of MicroSchema a new flexible XML-model for exchange of research documentation in Current Research Information Systems (CRIS) has been developed and proposed. A working XML-exchange model for metadata exchange between different CRIS and between with library systems and CRIS have been tested. A technical report describing the test case will be published summer 2002, the title of the report is: "Technical report of June 2002. Proposal for a flexible and extensible XML-model for exchange of research information by use of MicroSchema : Description of a working model for documentation produced by researchers".

## 5. Conclusion

A more flexible approach is needed to validate the exchange of data between different information systems. To solve this need, the concept of MicroSchema is introduced.

A new flexible and extensible XML-model for exchange of research information is proposed, using MicroSchema. The new XML-model has been tested against existing CRIS-systems, and data has been successfully imported into the model. The model has also with success been tested against ordinary library catalogue data.

## References

Biron, P.V. and Malhotra A., eds. 2001. XML *Schema Part 2: Datatypes*. The World Wide Consortium (W3C) http://www.w3c.org/TR/2001/REC-xmlschema-2-20010502

Bray, T.; Paoli, J.; Sperberg-McQueen, C.M. and Maler, E., eds. 2000. *Extensible Markup Language (XML) 1.0 (Second Edition)*. The World Wide Web Consortium (W3C). http://www.w3c.org/TR/2000/REC-xml-20001006

Fallside, D.C., eds. 2001. XML Schema Part 0:Primer. The World Wide Consortium (W3C) http://www.w3c.org/TR/2001/REC-xmlschema-0-20010502

ISO 8879:1986 *Information processing Text and office systems – Standard Generalized Markup Language (SGML)*

McClelland, M.; McArthur, D.; Giersch, S. and Geisler, G., 2002. Challengs for Service Providers When Importing metadata in Digital Libraries. In: *D-Lib Magazine*, 8 (4)

Thompson, H.S.; Beech D.; Maloney, M. and Mendelsohn, N., eds. 2001. *XML Schema Part 1: Structures*. The World Wide Consortium (W3C) http://www.w3c.org/TR/2001/REC-xmlschema-1-20010502

# Authors Index

## P



## Q



## R



## S



## T



## V



## W



## Z



## Ø