Knowledge Management and Organizational Learning 13

Isto Huvila
Lisa Andersson
Olle Sköld *Editors*

Perspectives on Paradata

Research and Practice of Documenting Process Knowledge

Springer

Knowledge Management and Organizational Learning

Volume 13

This series is introduced by the International Association for Knowledge Management (www.IAKM.net) with an aim to offer advanced peer-reviewed reference books to researchers, practitioners and students in the field of knowledge management in organizations. Both discussions of new theories and advances in the field, as well as reviews of the state-of-the art will be featured regularly. Particularly, the books will be open to these contributions: Reviews of the state-of-the art (i.e. syntheses of recent studies on a topic, classifications and discussions of theories, approaches and methods, etc.) that can both serve as a reference and allow opening new horizons Discussions on new theories and methods of scientific research in organisational knowledge management Critical reviews of empirical evidence and empirical validations of theories Contributions that build a bridge between the various disciplines and fields that converge towards knowledge management (i.e.: computer science, cognitive sciences, economics, other management fields, etc.) and propose the development of a common background of notions, concepts and scientific methods Surveys of new practical methods that can inspire practitioners and researchers in their applications of knowledge management methods in companies and public services.

This is a SCOPUS-indexed book series.

Isto Huvila • Lisa Andersson • Olle Sköld
Editors

Perspectives on Paradata

Research and Practice of Documenting Process Knowledge

Springer

Editors

Isto Huvila
Uppsala University
Uppsala, Sweden

Lisa Andersson
Uppsala University
Uppsala, Sweden

Olle Sköld
Uppsala University
Uppsala, Sweden

ISSN 2199-8663 ISSN 2199-8671 (electronic)
Knowledge Management and Organizational Learning
ISBN 978-3-031-53945-9 ISBN 978-3-031-53946-6 (eBook)
https://doi.org/10.1007/978-3-031-53946-6

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No 818210 as a part of the project CApturing Paradata for documenTing data creation and Use for the REsearch of the future (CAPTURE).

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

There has been a continuous and accelerated production of information since the birth of the Internet and the shift to digital formats. One of the outcomes of this growth has been the challenge of managing the vast amounts of information and being able to retrieve that which is both discoverable and provides useful information. Metadata, or data about data, provides basic information about data which makes it easier to find and use. Common elements of metadata include title, creator, date, subject or keyword, type of resource, and the rights and use. There are times, however, when additional contextual information would be helpful to users about the process and practices surrounding the creation of data. "Paradata" is an important concept that incorporates these ideas and provides the necessary contextual detail. Paradata derives from the Greek prefix "para" meaning alongside or beyond and refers to auxiliary data that surrounds a dataset. It can encompass a wide array of information, including timestamps, survey data such as response patterns and user interactions, and type of respondent that ensures the survey is reaching the target population.

Perspectives on Paradata: Research and Practice of Documenting Data Processes dives into the multifaceted realm of paradata, across a broad range of disciplines. Beginning with a comprehensive overview by the editors, the authors of this volume explore diverse applications, benefits, challenges, and future prospects of paradata in the context of their specific disciplinary approaches to contemporary research methodologies. By examining its role in enhancing data descriptions and illuminating details about the processes and practices involved, this book aims to shed light on the vital nature of paradata in the era of data-driven decision-making.

As noted in this volume, paradata can be applicable and useful across many research domains by enhancing the rigor and validity of data-driven endeavors. Paradata can provide evidence of the appropriateness of research procedures, increase trust in the results, and improve the (re)usability of earlier research results and data. It also allows researchers the opportunity to refine data collection instruments, identify and resolve errors or biases, and validate response patterns. Furthermore, paradata can improve the replicability and transparency of studies, leading to an enhanced analysis of findings.

In survey research, for example, paradata can aid in quality control, in response analysis, and in offering details about data sampling. Its applications can also extend to behavioral analysis, experimental design, and social sciences, sharing important insights into participant behavior and study practices. In addition to survey data, it can also include manual descriptions of procedures, automatically extracted and collected process data, and citations.

The editors note, however, that despite the many benefits of using paradata, there are ethical considerations surrounding privacy and consent that demand careful navigation. The technical complexities of paradata collection, storage, and analysis also require careful handling. Further research is needed to develop the necessary guidelines and parameters for how to best implement paradata while simultaneously considering the ethical considerations implicit in its practice.

This book offers a sweeping overview of paradata, highlighting its important role in research across many fields. It offers multiple understandings of what constitutes paradata and its definition, applications, benefits, challenges, and future trajectories. The book's editors, Isto Huvila, Lisa Andersson, and Olle Sköld, are all experts in their study of paradata and bring a wealth of knowledge and experience with paradata in their own areas of interest.

Isto Huvila is well recognized for his work in information and knowledge management and the considered and thoughtful approach he takes to examining theory and practice in information and knowledge work. He has worked in multiple domains including health information, social media, archives, libraries, museums, and knowledge management. He is the Director of project CAPTURE (CApturing Paradata for documentTing data creation and Use for the REsearch of the future), funded by the European Research Council.

Lisa Andersson has conducted research at the intersection of library and information science, digital humanities, and archaeology, with a focus on research data and documentation. She has examined how archaeology researchers cite methods in field observation documentation to better understand the types of paradata that can be found in their research. This leads her to discuss the challenges and opportunities in identifying paradata and the possibility of using paradata for assessing data reliability.

Olle Sköld has co-published in the area of paradata with Professors Huvila and Andersson and is a member of the CAPTURE team. His research focuses on the ALM field, knowledge organization and production, research data creation and use, and digital humanities.

The editors' expertise guides us through this book as they continue their quest to develop a deeper understanding of the challenging issues of information and knowledge management as manifested in the burgeoning field of paradata. By drawing from their own constellation of multiple domains, they skillfully bring together a wide range of disciplinary experts to describe the benefits and challenges of paradata in their respective fields. The result of these efforts is a book that explains and describes the value of paradata across multiple domains, facilitating knowledge sharing about the processes involved in research and allowing others to apply that detailed and contextualized knowledge to inform their own work. The future of paradata offers great promise for groundbreaking advancements.

Kent State University, Kent, OH, USA
October 2023

Kendra Albright

Acknowledgments

This volume is the result of wonderful group work and dedication by all the chapter authors. We as the editors are truly impressed by your dedication, your intellectual curiosity to explore an area sometimes far beyond your everyday research interests, and your open-mindedness to work in a truly interdisciplinary group of people, share ideas, exchange views, and give and receive feedback. The spirit and enthusiasm of the group was palpable during the 2 days we had the privilege to host you in Uppsala in the summer of 2022. In you, we have also been privileged to work with a group of colleagues who have kept deadlines and diligently delivered your texts to make this volume happen.

We would also like to thank all the anonymous peer reviewers for their reviews, constructive critique, and comments on the earlier versions of the chapters. Further, we extend our gratitude to our colleagues in the CApturing Paradata for documenTing data creation and Use for the REsearch of the future (CAPTURE) research project for productive discussions on paradata, feedback, and insights during the process: Zanna Friberg, Amalia Juneström, Jessica Kaiser, and Ying-Hsang Liu. We are also thankful to the advisory board of the project, including Costis Dallas, Melanie Feinberg, Jeremy Huggett, Eric Kansa, Costas Papadopoulos, and Sally Wyatt, for enlightening and intellectually rewarding discussions on paradata and documentation of processes. Finally, we would like to express our gratitude to professor Kendra Albright for agreeing to write a foreword to this volume.

This volume is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 818210) for the research project CApturing Paradata for documenTing data creation and Use for the REsearch of the future (CAPTURE).

Uppsala
October 2023

Isto Huvila
Lisa Andersson
Olle Sköld

Contents

About the Editors

Isto Huvila Ph.D., is professor in information studies at the Department of ALM at Uppsala University in Sweden. He received his MA degree in cultural history at the University of Turku in 2002 and a PhD degree in information studies at Åbo Akademi University (Turku, Finland) in 2006. Huvila was chairing the recently closed COST Action ARKWORK and is directing the ERC-funded research project CAPTURE. His primary areas of research include information and knowledge management, information work, knowledge organization, documentation, and social and participatory information practices.

Lisa Andersson Ph.D., works as a researcher at the Department of ALM at Uppsala University in Sweden. She received her MA degree in library and information science in 2011 and her doctoral degree in library and information science in 2017, both at the Department of ALM at Uppsala University in Sweden. Her research focuses on data and information management including research data and information management systems, knowledge organization and data descriptions, data publishing, and use. Andersson has published in library and information science journals but also in cross-disciplinary journals in the fields of archaeology and digital humanities.

Olle Sköld Ph.D., works as Senior Lecturer at the Department of ALM at Uppsala University in Sweden. He received his MA degree in archival studies in 2010 at Lund University and a doctoral degree in information studies in 2018 at Uppsala University. His research is characterized by a broad interest in the ALM field, knowledge organization and production, research data creation and use, and digital humanities. Sköld has published in information studies journals including the *Journal of the Association for Information Science and Technology*, *Journal of Documentation*, and *Information Research*.

An Introduction to Paradata

Lisa Andersson, Isto Huvila, and Olle Sköld

Abstract

To address the challenge of data process and practice descriptions, the aim of this volume is twofold. First, we propose the paradata concept as a method to think about and describe data, information, and knowledge processes and practices. Second, by inviting scholars and practitioners from a wide variety of disciplines, we explore how the paradata concept can be useful in and for information and knowledge management in a wide range of settings. The volume brings together scholars and practitioners from a broad range of subject areas, each offering a distinctive perspective on paradata in different contexts, highlighting diverse scenarios in which collection, extraction, and use of such data may prove constructive. The introductory chapter gives a brief history of the paradata term including insights into current research on the topic. Thereafter follows a concise contextualization of the study of paradata in the knowledge management tradition. The chapter is concluded by a guide to the volume's chapters.

1 Introduction

A pivotal question linked to data, information, and knowledge is what we need to know about data, information, and knowledge and how it has been managed in order to use it? Sometimes the answer is "very little." There is, for example, no harm in assuming that a temperature value is exact rather than rounded or not knowing the details of how a particular thermometer works when checking tomorrow's weather

L. Andersson · I. Huvila (✉) · O. Sköld
Uppsala University, Uppsala, Sweden
e-mail: lisa.andersson@ki.se; isto.huvila@abm.uu.se; olle.skold@abm.uu.se

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_1

forecast. At other times the situation may be the opposite; the precision of a value is crucial to its usefulness for a particular purpose, e.g., the precision of a temperature value for analyzing global warming. This is where paradata, i.e., information on the processes of data creation, curation, and earlier and intended use, can be pivotal for the usefulness of data, information, and knowledge.

Much of the intricacy of paradata lies in its ubiquity. Like Larry Prusak predicted the future of knowledge management, the knowledge of processes tends to be "thoroughly adopted—so much a natural part of how people organize work—that it eventually becomes invisible" when we engage in our daily pursuits (Prusak, 2001, p. 1006). However, as the temperature example shows, while simply accepting a temperature value can be sensible in many daily tasks, neglecting to account for the process can have major and possibly detrimental consequences in situations where precision is needed. To address the challenge of process and practice descriptions, the aim of this volume is twofold. First, we propose the paradata concept as a method to think about and describe process knowledge, particularly in relation to data, information, and knowledge processes and practices. Second, by inviting scholars and practitioners from a wide variety of disciplines, we explore how the paradata concept can be useful in and for information and knowledge management in a wide range of settings. By this exploration of process knowledge, how process information can support the management of information and knowledge, and of how process information and knowledge can be managed in a variety of contexts we hope to contribute to theoretical and practical advancement in the field of information and knowledge management.

Rather than starting with one particular definition of what paradata is, we embark on an exploration on what paradata can be, both in conceptual and practical sense. As a starting point of this journey, we borrow a definition of "data" from the information science scholar Christine L. Borgman stating that data refer to "entities used as evidence of phenomena for the purposes of research or scholarship" (Borgman, 2015, p. 29) and expand it to include other practices of creating knowledge or knowing also beyond scientific and scholarly pursuits. Further we start off with a working definition of the concept of paradata as "data on processes" following the seminal use of the term in survey research (see the chapter "Paradata in Surveys"). The working definition of paradata serves as a common ground throughout the chapters and creates a space for conceptual exploration of what paradata can be in different settings, the character of the processes they are meant to describe, the methods by which paradata are found or generated, what paradata can do or enable, and what needs to be considered when creating and using paradata in different settings and for different purposes. In the concluding discussion, we summarize and synthesize the various applications of the term as proposed by the 11 chapters.

The volume brings together scholars and practitioners from a broad range of disciplines, each offering a distinctive perspective on paradata in different contexts, highlighting diverse scenarios in which collection, extraction, and use of such data may prove constructive. The heterogeneity of the chapters included is by design; by juxtaposing different approaches to process data and engaging with them, we spur a wider discussion on the need to record—or reconstruct—data creation processes in order to uncover and make visible previously disregarded and invisible aspects of the creation, curation, and use of the many tangible and intangible resources and materials from which we gather data. Thus, while the theoretical and practical exploration of paradata is rooted in the information and knowledge management tradition, the volume's inquiry into process data and its management implications branches out into the plethora of fields with which the volume engages. To exemplify, the chapter "Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes" on paradata in the legal settings demonstrates how paradata can be necessary not only to describe how a piece of evidence came into being, but also to evaluate its evidentiary status in the legal process. As another example, the chapter "Paradata in Emergency Services Communications Systems" on research using emergency service communications data makes a note on two types of paradata, as something created in the course of research activities to make the research comprehensible and in parallel as something sought by researchers to understand the datasets they use. At its outset, rather than inquiring into paradata as a matter of concern for information and knowledge management only, the volume points to the multiple and varied uses of paradata in different scholarly and professional practices to provide insights into diversity of perspectives to paradata and to the intersections of the diverse approaches stemming from a multitude of frames of reference.

Through delving into paradata from the plethora of disciplinary perspectives included in this volume, the chapters forward the understanding of and relevance of paradata as a topic of interest for information and knowledge management. Building on the tradition of theorizing and developing metadata to serve resource description purposes, the inquiry into paradata prompts several leaps forward into new theoretical and practical challenges to engage with. One of the key issues in the earlier paradata-related literature, as well as throughout the chapters of this volume, is how paradata can help to make cognizable and draw attention to doings that underpin different types of information and knowledge. In this sense, the concept points directly to one of the thorniest and most prominent knowledge management problems—that of how to deal with tacit and implicit aspects of knowledge and bring them together with explicit and inscribed forms of information and data (Polanyi, 1967). At the same time, paradata also meanders somewhere in between the two major perspectives to knowledge management as either a social and organizational issue of mediating and translating knowledge between people or a technical issue of inscribing and managing knowledge in, often, technical systems (Handzic, 2004). Depending on how paradata is conceptualized, it can feature as a translational device from either of the two perspectives to bridge the gap in between. Moreover, because paradata often serves to enable a particular data reuse purpose, paradata opens up for theorizing and developing practical solutions to span the spectrum between general, definitive resource description and process information created for situated and time-specific data reuse needs. As with the initial temperature example, an existing set of temperature data can at a later point in time be enriched with thermometer configuration data, if and when the data is about to be aggregated with other temperature datasets for a more comprehensive analysis.

On a practical note, besides informing practice in the disciplines represented in the individual chapters, the volume will be useful for information and knowledge management professionals seeking to develop paradata documentation in various practices. Also professionals working specifically with data management such as data stewards, data curators, data managers, and data librarians will find the various chapters useful both for drilling deeper into process descriptions in particular disciplinary practices and for understanding the breath and differences between various fields of research and practice. Even information architects and system developers designing data repositories and services for data discovery and reuse will find the volume useful for similar purposes. For researchers and master's and doctoral students the volume will offer a comprehensive overview of a fast-growing field of study with significant and broad relevance for data, information, and knowledge management. Moreover, anyone creating or dealing with data in their daily work will find the volume as food for thought when reflecting on how to describe data processing in a meaningful and efficient way.

This introductory chapter will in the following give a brief history of the paradata term including insight into current research on the topic. Thereafter follows a concise contextualization of the study of paradata in the knowledge management tradition. The chapter is concluded by a guide to the volume's chapters.

2 The Origins and Status of the Term "Paradata"

The "para-" prefix (Definition of PARA, n.d.) means beside or alongside of, and thus would make paradata mean information alongside of data. Confusingly, the more established term "metadata" can also mean data about data (Pomerantz, 2015). In practice though, metadata often refers to a condensed and structured description of a resource, commonly guided by a standard agreed upon by some form of community. A well-known example of a metadata standard is the Dublin Core standard developed to describe networked electronic information objects, encompassing categories like "Title," "Type," "Subject," etc. (DCMI Metadata Terms, n.d.). In order to understand what paradata can bring to the table in terms of data descriptions, it is vital to understand what metadata traditionally does not provide data creators and users a chance to discern and do.

Paradata, as applied for example in statistics to describe survey data, implies a description of the survey process (Couper, 2000; for further description, see the chapter "Paradata in Surveys," Paradata in survey research). Similarly, in heritage studies, paradata has been used to denote information on data creation and processing (Denard, 2012; see also the chapter "A Leap of Faith: Revisiting Paradata in 3D Scholarship," dismantling the black boxes of 3D scholarship). Even if there is obvious overlap between what could count as metadata and paradata both in theory and practice, in the light of the earlier literature, the paradata notion has a unique and developing conceptual space. Paradata points to a need to document data processing beyond what is traditionally captured in structured metadata. As a phenomenon, paradata differs from metadata qualitatively in that it seeks to cover the creation and processing of data and quantitatively in that it encompasses more detailed information about data than traditional metadata. Also, according to how it has been conceptualized, the engagement with paradata disrupts established data description standards and practices as data processing information to a large extent is unstructured and not codified. As a positive consequence of that paradata is not (yet) formalized to the same extent as metadata, the previous research and the chapters of this volume show how paradata also unfolds as a potentially powerful device with which to think and explore matters tied to how doings and processes can be documented and understood across domains.

Parallel to paradata, there are other terms partly overlapping in meaning. “Provenance” is a concept common in the archival sphere to describe both creation and curation of informational objects, specifically records. It can encompass both the agents involved and the actions they take, and be useful for acquiring, arranging, retrieving, and appraising records (Sköld et al., 2022). Consequently, and as this edited volume illustrates, scholars and practitioners in information science and knowledge management are not alone in grappling with the challenge to share and understand information on why a data source came to be the way it is. Yet, as of today, the character of and need for paradata remains open for exploration in most fields of data creation. Likewise, the question of its relation to metadata and the challenge of incorporating processing information into structured resource descriptions is in the stage of experimentation. There are well-defined models that cover specific types of paradata, like the CIDOC-CRMdig (Doerr et al., n.d.) standard for documenting the steps and methods of producing digital objects, PROV specification for modeling data provenance (PROV-Overview: An Overview of the PROV Family of Documents, 2013), and Common Workflow Language for inscribing computational analytical workflows (Amstutz et al., 2016). While these models meet parts of users’ paradata needs, a recent analysis of the use of the paradata concept in archaeology and heritage studies points to a number of uncertainties pertaining to the paradata concept that require further study to be clarified. These uncertainties include the challenge of determining the required types and amounts of paradata, considerations of paradata users, and what kind of transparency paradata can be expected to facilitate (Sköld et al., 2022).

The next section takes a closer look at the exploration of paradata in the field of knowledge management and how the emerging interest in paradata both taps into traditional trajectories and opens up new paths for knowledge management theory and practice.

3 Paradata in Information and Knowledge Management

3.1 Approaching Paradata

The exploration of paradata as seen in this volume and in multiple research literatures has apparent affinities with information and knowledge management, a field of research and practice that concerns the creation, capturing, organization, access, and use of sources of information and knowledge. In this volume, we refer to information and knowledge management as the broad field of research and practice that comprises and is termed in the literature varyingly as knowledge management or information and knowledge management with a focus on managing knowledge, information and data, records, and collections. This framing acknowledges the diversity of the field both regarding differences in research traditions and their analytical focus, including the differences in referring to knowledge management, organizational learning, information management, data management, and records management (Pun & Nathai-Balkissoon, 2011; Schlögl, 2005). However, for us in this volume, similarly to how paradata has been discussed in relation to metadata so far, the boundaries between "information" and "knowledge" are definitional rather than strict. According to both the widely adopted and criticized data–information–knowledge hierarchy, information is commonly understood as a message, and knowledge as information that is contextualized in the minds or practices of knowing individuals or organizations (Davenport, 1998). While previously much of the information and knowledge management literature has focused on one of the two or both concepts, interestingly from the perspective of paradata, lately also the terms "data" and "data management" have been related to and included in the wider field of knowledge management (Dalkir, 2023).

Even if much of the knowledge management literature relates to business contexts referring to the management of knowledge for the efficacy and profit of the business organization (Bolisani & Bratianu, 2018), it has been used in a wider sense as denoting the dealing with information or knowledge as a means to an end—whether it be personal or professional, for leisure or profit. Knowledge management is a thriving topic in the context of library management (Shropshire et al., 2020), and there is a growing corpus of work on personal everyday life information and knowledge management (Dinneen & Julien, 2020; Pauleen & Gorman, 2011; Swigon, 2011). An earlier book in the International Association for Knowledge Management series *Knowledge Management and Organizational Learning*—where this volume is appearing—focused on knowledge management in and in relation to arts and humanities (Handzic & Carlucci, 2019). By inquiring into paradata, we are continuing this interdisciplinary push toward expanding the horizons of information and knowledge management research and practice. Knowledge management research generally attempts to theorize beyond the specific case or cases, using empirical evidence to generate models with a more general explanatory reach. Yet, at the same time it recognizes the cultural, social, and cognitive aspects of and influences on managing knowledge and information. Much similarly to how information and knowledge management has many practical and theoretical applications in arts and humanities, and arts and humanities perspectives can inform the development of information and knowledge management theory and practice, the multidisciplinary engagement with paradata as put forth by this volume has implications in the affected disciplines and on paradata as an information and knowledge management concept.

Just like how information and knowledge management commonly concerns the system level management of information and knowledge also paradata unfolds as a potentially comparably systemic concept. "System" here refers to technical systems for recording, storing, and sharing information, but also the organizational and social system governing information. The knowledge in question can be both that can be codified and stored in a technical system and the lived knowledge that people know by experience and act on by default. Similarly to a part of the previous and ongoing paradata research, along the lines of the technical strand of information and knowledge management, to reach its goal of efficient information and knowledge sharing, much of the research effort has been put into understanding how knowledge can be codified, shared, and retrieved by support from technical systems. Yet, interdisciplinary work on both paradata and information and knowledge management alike also takes interest in the knowledge sharing taking place beyond technical systems—both to understand the flaws of technical systems in order to improve their functionality and to understand what knowledge cannot be codified and transmitted via a technical system. A now classic study of customer support service operators shows the supremacy of sitting next to the most knowledgeable co-worker, rather than using the knowledge management system provided (Orr, 2016). Correspondingly, the human and the lived experience as sources of paradata should not be overlooked, as evinced by multiple contributions to this volume like the chapter "Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites" on paradata derived from analytical narration of a fieldwork experience and the chapter "Towards Embodied Paradata. A Diffractive Art/Archaeology Approach" on what we can learn about processes from the bodily and embodied information.

3.2 Data Descriptions and Knowledge Management

Somewhat counterintuitively, even if the work on taxonomies and development of knowledge management systems have acknowledged and addressed to a degree the significance of organizing resources, the information and knowledge management discipline has never put the description and organization of knowledge in its immediate focus. Therefore, it is perhaps unsurprising that the conceptual work on process descriptions and paradata stems from other disciplinary contexts. Nevertheless, how sources are described and organized is pivotal to their findability and usability in any system also from the information and knowledge management perspective.

In contrast to the relative lack of emphasis of descriptions and the work of describing and organizing, the description and organization of sources is the key focus of the neighboring discipline of knowledge organization (KO) (Smiraglia, 2014). Knowledge organization research deals with metadata as a theoretical construct and practical implementation to make knowledge, information, and data findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). The idea of knowledge, information, and data reusability is the intersection where the interests of knowledge organization and information and knowledge management merge. For a knowledge management system to meet its objectives, the knowledge within needs to be not only findable but also reusable, and reusability presupposes some extent of understanding of what is being made findable and reusable both in terms of their whatness and their processual origins.

But resource description has never been easy. Both knowledge organization and knowledge management have since long acknowledged the complexity of *organizational* as opposed to *individual* knowledge, the *cultural* and *social* contexts of knowledge, and the *tacit* versus the *explicit* components of knowledge. However, the challenges of describing and managing resources differ depending on the type of resource in focus of the description and management. The resource might be as different as a piece of "know-how," the solution to a common problem traded between software developers writing code, or a static fact about the head count of a customer organization. However, when the resource is "data"—be it a continuous data lake or a bounded dataset, expected to be reusable and have sustained usability over time, the resource presents a new set of descriptional challenges connected to how the data came into being, was organized, and has been managed over time. This is where paradata has the potential to support knowledge management.

While traditional metadata aims at describing sources by assigning attributes like "language" or "date issued," and sometimes even gives insight into the history of a source by stating "provenance" indicating changes is ownership or custody (DCMI Metadata Terms, n.d.), the rationale behind expanding traditional metadata with additional information on the origin and processing of data is that users likely will need to know more than core metadata in order to understand a data resource to the extent that it becomes usable (Börjesson et al., 2022). Thus, to meet the knowledge management goals about efficient retrieval and reuse of knowledge in the form of data, the data resources need to be sufficiently well-described—often beyond what is achieved by traditional metadata.

Thus, to understand paradata as a rising topic of practical importance and information and knowledge management research interest we need to look at the resource in focus of the management efforts. From a narrow perspective, organizations aim to manage, reuse, and capitalize off of their data assets, and from a broader perspective, societies need to manage their data in sustainable ways to prevent waste of public resources and build critical knowledge, e.g., about demographics and public health, over time. Furthermore, from a research perspective—as the datasets become larger and the methods more advanced—researchers need refined ways of describing their methods to maintain methodological transparency and produce reliable results, as is apparent for example in the case of genetic epidemiology in the chapter "Making Research Code Useful Paradata." The following final section of this introduction gives a thematic overview of the chapters offering insights into their contributions to the emerging field of paradata research.

4 Thematic Overview

While this volume collects contributions concerning paradata from a range of fields, it does by no means offer an exhaustive overview of paradata types and paradata use cases. We hope and believe that giving a more comprehensive overview can be a task for further research. The selection of fields and practices covered in this volume is based on us, the editors, identifying contexts and practices where process descriptions are needed and, in some instances, already explored and practiced. Some, like survey research (chapter "Paradata in Surveys") and heritage visualization (chapter "A Leap of Faith: Revisiting Paradata in 3D Scholarship"), had a given place, as those are fields where the paradata term has been used already for several decades. Others, like archival science (chapters "Mapping Accessions to Repositories Data: A Case Study in Paradata" and "The Role of Paradata in Algorithmic Accountability") and legal analysis (chapter "Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes"), were identified as domains where process descriptions are needed to complete core tasks of the fields. We reached out to scholars in these areas asking them to describe and analyze methods for process descriptions within their field or area of expertise and to use the notion of paradata to discuss the process description and documentation. As a result, the volume combines contributions from authors familiar with using the paradata concept already before writing their chapters and contributions from authors thinking with and describing process descriptions as paradata for the very first time. While we provided the starting point for a conceptually driven exploration, the analyses and reflections are wholly attributable to the chapter authors.

As a result of our highly purposive sampling method, you will meet authors from a wide range of fields and practices, bringing their disciplinary styles and stylistic preferences to the table. While the sample of disciplines and authors is by no means systematic or meant to be such, all chapter authors do, to some degree, engage either explicitly or implicitly in information and knowledge management in their writing—although many have never been aware of or related their work to information and knowledge management theory and research previously. Sometimes the links are obvious, sometimes less so. In the latter case, rather than forcing the chapters to take steps outside of their home domains, we have made an effort to reflect on the links in the final chapter of this volume (Huvila, Andersson, & Sköld, this volume). Moreover, the volume covers chapters on processing descriptions in relation to public as well as private data, research, and governance data. The purposes for storing and processing data vary accordingly, from commercial purposes to heritage preservation.

Naturally, the quest of making process and practice information and knowledge more explicit and more tangible speaks to contemporary, sometimes highly ideological, discourses of openness and transparency. These are familiar from the open science movement (Open Innovation, Open Science, Open to the World | Shaping Europe's Digital Future, 2016), efforts to make the use of personal data more transparent, such as the EU's GDPR (General Data Protection Regulation (GDPR), n.d.), and call for insight into the algorithmic processes shaping our lives (see for example the chapter "The Role of Paradata in Algorithmic Accountability"). Paradata can thus be drawn upon to promote positive values such as accountability, trust, and sustainability in information and knowledge management processes. At the same time, the notion of paradata raises the challenging questions of to what extent intellectual and social as well as technical processes can be described and understood. The task of creating understandable process descriptions challenges our human creativity, as well as the modes of communication we have available, and the forms of documentation we use to inscribe and transmit the communication. This way, the notion of paradata is both an expression of a quest to describe and manage process information and a humbling reminder of how difficult it is to capture process knowledge.

As already noted, the chapters of the volume span between paradata created in research, either as a support for the research process as in the case with computational code in the chapter "Making Research Code Useful Paradata" or as more of a by-product of a study, as exemplified by the chapter on surveys, the chapter "Paradata in Surveys." Also, we see a breadth in the reasons for documenting paradata. Process documentation for transparency of processes for current and future audiences is one motivation, apparent for example in the chapter "A Leap of Faith: Revisiting Paradata in 3D Scholarship" on heritage visualizations and the chapter "Paradata for Digitization Processes and Digital Scholarly Editions" on digitalization processes. Paradata as a direct input into the management, in the sense of control or change of processes, is another motivation, standing out for example in the chapter "Paradata in Emergency Services Communications Systems" on emergency service communication and the chapter "Adding Paradata About Records Processes via Information Control Plans" on public-sector processes.

The volume starts by a chapter by the statistician Patrick Oliver Schenk and the archaeologist Simone Reuß, introducing the field of survey studies where the paradata is a widely adapted concept. The chapter probes deeper into the definition of paradata by comparing the term to the related terms auxiliary data, contextual data, and metadata, and gives ample examples of the variety of methods for collecting paradata. The chapter "Making Research Code Useful Paradata," by the computational biologist Richèl J.C. Bilderbeek, delves into a specific type of paradata, namely computer code in computational research for the purpose of reproducibility. Using the case of genetic epidemiology, Bilderbeek proposes how to improve code to make it serve as paradata. The chapter "A Leap of Faith: Revisiting Paradata in 3D Scholarship," by the digital humanities scholar Costas Papadopoulos, brings us into the practice of 3D (re)construction where the formats for paradata are far more composite. Papadopoulos highlights the variable and dialectic processes of fieldwork and points to the perceptual, physiological, and technical factors that need to be accounted for to understand 3D (re)constructions.

The chapter "Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites," by the archival studies scholar Sarah A. Buchanan and the archaeologist Theresa Huntsman, also concerns paradata from fieldwork, but focusing on the lived experience as paradata. In their framing, paradata is the collaborative analytical narration, led by a data archivist to support the contextual integrity of the data collected, and to surface the pedagogical goals of each project. The artist and art scholar Ian Dawson and the archaeologist and computer scientist Paul Reilly similarly emphasize the involvement of the human subject in the creation of paradata in the chapter "Towards Embodied Paradata. A Diffractive Art/Archaeology Approach." Dawson and Reilly propose the notion of "embodied paradata" as a way to understand how the worker and their tools, their bodily practices of making or uncovering knowledge, make up paradata.

The chapters "Mapping Accessions to Repositories Data: A Case Study in Paradata," "Paradata for Digitization Processes and Digital Scholarly Editions," and "Reconstructing Provenance in Long-Lived Data Systems: The Challenge of Paradata Capture in Memory Institution Collection Databases" put the paradata concept to work for the purpose of understanding the composition of collections, editions, and databases. In the chapter "Mapping Accessions to Repositories Data: A Case Study in Paradata," the historian of science Kevin Matthew Jones and the archivist Jenny Bunn explore how paradata can shed light on the choices and assumptions made by archivists in accession processes. In the chapter "Paradata for Digitization Processes and Digital Scholarly Editions," the information science scholar Wout Dillen discusses how digitized collections often lack the information of how the digitization was made, and thereby leaving the user of the digital reproductions without the necessary clues to understand the quality of the reproduction. Paradata on the digitization process would, in this case, directly improve research validity. The chapter "Reconstructing Provenance in Long-Lived Data Systems: The Challenge of Paradata Capture in Memory Institution Collection Databases," by information science scholars Alexandria Rayburn and Andrea Thomer, explores how databases' histories and maintenance can be documented by means of visualizations and thereby made accessible to users.

A number of chapters also apply the paradata concept in situations beyond research and collections development and bring insight into various processes highly definitional of people's daily lives. The multidisciplinary team behind the chapter "Paradata in Emergency Services Communications Systems" explores how paradata in government archives could support the modeling and simulations needed to try out new technologies for emergency service communications. The information science scholars Ciaran B. Trace and James A. Hodges delve into the role of paradata for algorithmic transparency and explainability of algorithms and algorithmic systems in the chapter "The Role of Paradata in Algorithmic Accountability." The chapter "Adding Paradata About Records Processes via Information Control Plans" explores how automatically assigned paradata could contribute to information control in public administration. The archival scholars Saara Packalén and Pekka Henttonen discuss how information on the processes from which records originate may further the understanding of the records kept. Lastly, the law scholar Lena Enqvist, similarly to Packalén and Henttonen, explores how paradata could increase transparency in the public sector. Enqvist investigates how paradata could make public authorities' use of technology in decision making more visible and auditable from a legal perspective.

The concluding chapter draws together insights from the discipline-specific chapters to contrast and synthesize the diverse approaches to how paradata is conceptualized and used. Further, we proceed to three topics of discussion emerging from the synthesizing analysis: how paradata is done in practice, the implications of paradata for the theory and practice of information and knowledge management, and the ethics of paradata. The chapter is concluded by brief remarks on future directions of paradata research and practice. By this exploration of paradata we do, together with the chapter authors in this volume, hope to promote the paradata concept as a valuable contribution to the knowledge management toolbox.

References

Amstutz, P., Crusoe, M. R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., Scales, M., Soiland-Reyes, S., & Stojanovic, L. (2016). *Common workflow language, v1.0* (5921760 Bytes) [Data set]. figshare. https://doi.org/10.6084/M9.FIGSHARE.3115156.V2

Bolisani, E., & Bratianu, C. (2018). *Emergent knowledge strategies* (Vol. 4). Springer. https://doi.org/10.1007/978-3-319-60657-6

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Börjesson, L., Huvila, I., & Sköld, O. (2022). Information needs on research data creation. *Information Research, 27*(special issue), isic2208. https://doi.org/10.47989/irisic2208

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review, 18*(4), 384–396. https://doi.org/10.1177/089443930001800402

Dalkir, K. (2023). *Knowledge management in theory and practice*. The MIT Press. https://mitpress.mit.edu/9780262048125/knowledge-management-in-theory-and-practice/

Davenport, T. H. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business School.

DCMI Metadata Terms. (n.d.). Retrieved May 26, 2023, from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Definition of PARA. (n.d.). Retrieved August 12, 2022, from https://www.merriam-webster.com/dictionary/para

Denard, H. (2012). *Paradata and transparency in virtual heritage*. Taylor & Francis.

Dinneen, J. D., & Julien, C.-A. (2020). The ubiquitous digital file: A review of file management research. *Journal of the Association for Information Science and Technology, 71*, E1–E32. https://doi.org/10.1002/asi.24222

Doerr, M., Stead, S., & Theodoridou, M. (n.d.). *CRMdig v4.0*. FORTH. Retrieved May 26, 2023, from https://www.cidoc-crm.org/crmdig/sites/default/files/CRMdigv4.0.pdf

General Data Protection Regulation (GDPR). (n.d.). Retrieved July 5, 2023, from https://gdprinfo.eu/

Handzic, M. (2004). *Knowledge management: Through the technology glass*. World Scientific.

Handzic, M., & Carlucci, D. (Eds.). (2019). *Knowledge management, arts and humanities*. Springer. https://doi.org/10.1007/978-3-030-10922-6

Open innovation, open science, open to the world | Shaping Europe's digital future. (2016, June 16). https://digital-strategy.ec.europa.eu/en/library/open-innovation-open-science-open-world

Orr, J. E. (2016). *Talking about machines: An ethnography of a modern job*. Cornell University Press. http://ebookcentral.proquest.com/lib/uu/detail.action?docID=4742043

Pauleen, D., & Gorman, G. (Eds.). (2011). *Personal knowledge management: Individual, organizational and social perspectives*. Gower.

Polanyi, M. (1967). *The tacit dimension*. Routledge.

Pomerantz, J. (2015). *Metadata*. MIT Press. http://ebookcentral.proquest.com/lib/uu/detail.action?docID=4397948

PROV-Overview: An Overview of the PROV Family of Documents. (2013). http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Prusak, L. (2001). Where did knowledge management come from? *IBM Systems Journal, 40*(4), 1002–1007. https://doi.org/10.1147/sj.404.01002

Pun, K. F., & Nathai-Balkissoon, M. (2011). Integrating knowledge management into organisational learning: A review of concepts and models. *Learning Organization, 18*(3), 203–223.

Schlögl, C. (2005). Information and knowledge management: Dimensions and approaches. *Information Research, 10*(4).

Shropshire, S., Semenza, J. L., & Koury, R. (2020). Knowledge management in practice in academic libraries. *IFLA Journal, 46*(1), 25–33. https://doi.org/10.1177/0340035219878865

Sköld, O., Börjesson, L., & Huvila, I. (2022). Interrogating paradata. In: *Information research. Proceedings of the 11th international conference on conceptions of library and information science, Oslo Metropolitan University, May 29–June 1, 2022* (Vol. 27, special issue, paper colis2206). https://doi.org/10.47989/ircolis2206

Smiraglia, R. P. (2014). *The elements of knowledge organization* (1st ed.). Springer. https://doi.org/10.1007/978-3-319-09357-4

Swigon, M. (2011). Information limits: Definition, typology and types. *Aslib Proceedings, 63*(4), 364–379.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. -W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18

Lisa Andersson Ph.D., works as a researcher at the Department of ALM at Uppsala University in Sweden. She received her MA degree in library and information science in 2011, and her doctoral degree in library and information science in 2017, both at the Department of ALM at Uppsala University in Sweden. Her research focuses on data and information management including research data and information management systems, knowledge organization and data descriptions, data publishing, and use. Andersson has published in library and information science journals but also in cross-disciplinary journals from the fields of archaeology and digital humanities.

Isto Huvila Ph.D., is professor in information studies at the Department of ALM at Uppsala University in Sweden. He received an MA in cultural history at the University of Turku in 2002 and a Ph.D. degree in information studies at Åbo Akademi University (Turku, Finland) in 2006. Huvila was chairing the recently closed COST Action ARKWORK and is directing the ERC-funded research project CAPTURE. His primary areas of research include information and knowledge management, information work, knowledge organization, documentation, and social and participatory information practices.

Olle Sköld Ph.D., works as Senior Lecturer at the Department of ALM, Uppsala University, Sweden. He received his MA degree in archival studies in 2010 at Lund University and a doctoral degree in information studies in 2018 at Uppsala University. His research is characterized by a broad interest in the ALM field, knowledge organization and production, research data creation and use, and digital humanities. Sköld has published in information studies journals including the *Journal of the Association for Information Science and Technology*, *Journal of Documentation*, and *Information Research*.

Paradata in Surveys

Patrick Oliver Schenk and Simone Reuß

Abstract

Paradata are widely used in conjunction with surveys, from predicting behavior for targeted interventions, monitoring data quality and interviewer performance, to understanding and correcting biases in the data. We define survey paradata broadly: as nonsubstantive data that relate to the survey and its processes in at least one of three ways—they are produced by survey processes, describe them, or are used to manage and evaluate them. They typically would not exist without the survey. They may be automatically produced (e.g., keystrokes), actively collected (e.g., interviewer observations), or constructed later on (e.g., when a human labeler rates respondent–interviewer rapport by listening to recordings).

First, we review other data types (auxiliary, contextual, and metadata) because their overlaps with paradata can make it difficult to grasp paradata precisely. We discuss paradata definitions, including their weaknesses, arriving at our definition.

Second, we offer an overview of our field's practice and literature: paradata examples, heterogeneity across paradata types and design options, applications, and challenges.

With paradata a somewhat mature concept in our field, survey methodology, we hope to provide a stimulating, broad introduction to practice and literature in our field, accessible to anyone irrespective of professional background. We

P. O. Schenk (✉)
Department of Statistics, Ludwig-Maximilians-University, Munich, Germany
e-mail: patrick.schenk@stat.uni-muenchen.de; p.o.s.on.stats@gmail.com

S. Reuß
Institut für Vor- und Frühgeschichtliche Archäologie, Ludwig-Maximilians-University, Munich, Germany
e-mail: reuss@vfpa.fak12.uni-muenchen.de

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_2

hope that this chapter provides a valuable backdrop for the conceptualizations of paradata in other disciplines, as presented in this volume.

1 Introduction

Surveys and Survey Methodology A survey is a systematic, standardized data collection effort that proceeds mainly by asking questions and recording the responses of the so-called survey participants or respondents.[1] Sometimes, data are not collected about individuals but about, e.g., households or firms—yet, ultimately it is still humans providing answers. Surveys can be carried out by interviewers, be self-administered by the respondents, or take on a hybrid form in which interviewers are present but, after handing over a tablet or paper questionnaire, are inactive unless needed. More specifically, the survey 'mode' is about how a survey is conducted: interviews in person or by telephone, web-based surveys and mail questionnaires (both self-administered), or even multiple modes for one survey across regions or persons, or time.[2]

Surveys have important benefits. They can cover a broad range of topics and can also probe for very detailed information. They can be tailored to the specific research questions for which a survey is conducted. Also, a person's attitudes, past experiences, or other information not recorded anywhere may be best or even only available by asking them. Finally, surveys benefit from decades of methodological research (see below).[3] Thus, academic research, government agencies, public opinion research and polling, and the private sector will continue to rely on surveys.

This chapter looks at paradata specifically from the perspective of survey methodology. Most survey methodologists originally come from the social sciences, psychology, and statistics. However, the field has its own terminology, goals, challenges, and thinking. Thus, how we regard and use paradata is hopefully best understood against the background context we provide next.

Surveys are conducted to answer substantive research questions,[4] mainly through statistical analysis of the collected data. Survey methodology studies the design,

[1] Our introduction relies on Groves et al. (2009). Note that surveys are distinct from the less structured qualitative methods that also use interviewing: of experts, focus groups, etc.

[2] In this chapter, we will repeatedly refer to the interviewers. Such statements, of course, only apply to survey modes that feature them. We omit this constant reminder.

[3] New data sources such as Big Data have gained prominence due to their often low cost and large volume. However, they rarely offer the breadth and level of detail of a survey, researchers typically have little influence over and information about what and how data are captured, and data quality can be very problematic. Survey data and these other data sources can, of course, be complements (Japec et al., 2015, p. 873 and Couper, 2017a, p. 134f.) and help to improve one another's methodology (Hill et al., 2021). Surveys are here to stay.

[4] 'Substantive' in the sense of the substantive/empirical sciences (studying real-world phenomena), as opposed to the methodological/formal sciences (doing methods research).

execution, and monitoring of surveys as well as the statistical analysis of survey data (Groves et al., 2009, ch. 1.4): what are sources of problems, how can they be measured, and what are methods to address them? Besides the quality of data and of data analysis, costs are a consideration.[5] A ‘survey organization’ (the entity conducting the survey) needs to be profitable. Costs also relate quality, quantity, and scope of a survey as well as the number of surveys that can be afforded, thereby influencing the number of research questions that can be answered. Survey methodologists care about costs, besides due to these real-world constraints, also because they imply data quality trade-offs.

Survey methodologists often think about the quality of data and of data analysis along two dimensions: the survey participants (*representation*) and the responses (*measurement*). Low participation rates, an increasingly grave problem for our field, increase the costs to achieve a fixed number of respondents. *Representation*, meanwhile, is about differences between respondents and nonrespondents: the less the respondents are representative of the population about which a researcher[6] wishes to draw inferences, the more likely the data (analyses) are biased. Representation problems can occur at every step. First, there is rarely a complete list of all units in a population. The ‘sampling frame’ is the incomplete list that *is* available, from external sources or through construction, e.g., a ‘lister’ walking a neighborhood to collect addresses. ‘Coverage’ is about how representative the sampling frame is of the intended target population. Second, only a subset of the units from the sampling frame is actually selected for possible inclusion in the survey (‘sample’). Third, who from the selection ends up in the data is then determined by who is not successfully contacted and then by who decides to not participate (‘unit nonresponse’).

Imagine the eventual survey data as a rectangular, tabular data sheet: each row corresponds to exactly one respondent, and each column corresponds to exactly one survey question. Representation is then about how representative the rows are. Meanwhile, *measurement* considers each cell and asks whether and how the value in this cell deviates from the true value that one intended to capture.[7] Such errors can, again, occur at every step from planning to data analysis. First, researchers start from ideas or conceptions (‘constructs’) about the “elements of information” they seek (Groves et al., 2009, ch. 2.2.1, 2.3.1). Some constructs have decent, objective counterparts in the real world, but many are ‘latent’ (‘unobservable’, not directly measurable). ‘Validity’ is about the degree to which the concept in the researcher’s mind matches how respondents understand the corresponding survey question. The precise wording of a question, the question order, and other design factors can

[5] The other, “fitness-for-use” quality dimensions of surveys are important per se but are not at the forefront of survey *methodology*: relevance, credibility, and timeliness (Groves et al., 2009, ch. 2.6).

[6] We use ‘researchers’ as a shorthand. Their substantive research questions are to be answered by the eventual analysis of the survey data, and they thus also influence what the survey is about.

[7] One may also look at a column as a whole: is there a difference between what this survey question should capture and the values that one got?

affect how participants interpret a question. Second, a response can deviate from the truth because of, e.g., recall error, low motivation, but also interviewer effects on, e.g., sensitive questions. A participant may also be unwilling or unable to respond to a particular survey 'item'[8] at all: 'item nonresponse'. Third, an initial or raw response may be processed or edited: the respondent might change their initial answer later or the interviewer might edit it. If only categorical answers are permitted, the initial response must also be mapped to the given categories. Sometimes, raw information is processed later on by 'labelers' (also called coders or annotators): e.g., coding open-text responses into categories or rating respondent behavior based on recordings.

Motivation We discuss other data types and definitions of paradata in Sect. 2. In short, our own definition (see 2.2) is that paradata are data that are themselves not the survey's substantive data, that would typically not exist without the particular survey, that can be automatically produced, actively collected, or constructed later on, and that relate to the survey and its processes in one of three ways: they are produced by the survey processes, often as by-products, they describe the survey processes, or they are used to manage and evaluate the survey process(es).

There are many different examples of survey paradata (Sect. 3), with much heterogeneity between them but also ample scope for designing them as needed (Sect. 4). The allure to survey methodologists comes from the goals and challenges laid out above: paradata can be employed to recognize problems in the survey data, to correct for them in the statistical analysis, and to monitor problems in near real-time or even to predict them which is the basis for interventions (Sect. 5). The hope is that paradata capture information about the processes that produce the survey data that would otherwise not be available. Some paradata types have been used in our field long before Couper (1998) first coined the term 'paradata'.[9] As survey methodologists, we touch on direct uses of paradata in substantive research only very briefly. However, while paradata may help with problems in the substantive data, they themselves also face and pose challenges (Sects. 6 and 7). Still, the message of our broad overview is a positive one: there are low-cost paradata types that offer a great starting point.

2 Paradata and Other Data

In Sect. 2.1, we review four data types that appear in conjunction with 'paradata' in the literature. A key takeaway will be that there are overlaps between them and paradata, as simplified and visualized in Fig. 1. In Sect. 2.2, we discuss paradata: definitions, the relation to 'process' and process data, and our own, broad definition.

[8] This is the more general and technical term than 'question' as surveys also contain, e.g., prompts that are not phrased as questions.

[9] See Couper (2017b, p. 6 and 11) who also reflects on the evolution of the concept 'paradata'.

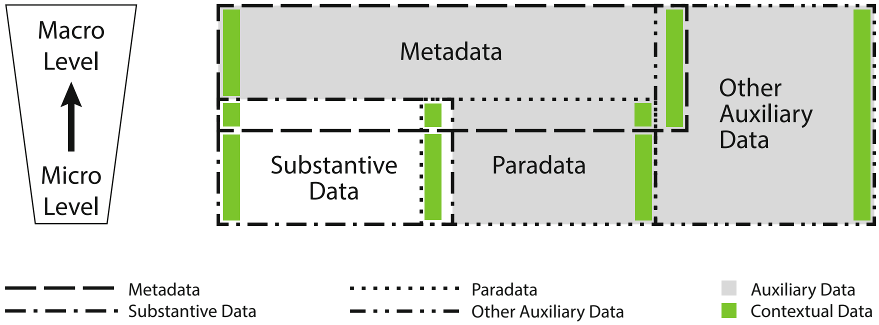


Fig. 1 Relation among data types. Sizes and overlaps are not proportional

2.1 Substantive Data, Metadata, Auxiliary Data, Contextual Data

Substantive Data[4] are “what surveys are designed to collect or produce” (Couper, 2017b, p. 4): they correspond largely to the participants’ survey responses but also include, e.g., samples and measurements taken by respondents, interviewers, or sensors (Groves et al., 2009, ch. 2.2.2; Keusch et al., 2024). Unfortunately, ‘data’ are used as both a synonym for substantive data and an umbrella term for *all* data types (i.e., including substantive data, paradata, metadata, and so on).

Metadata are, nowadays, “any descriptive information about some object(s) of interest” (NAS, 2022, p. 96). Thus, survey metadata are information about the survey, its components, and the produced data—“the core of [survey] documentation” (Kreuter, 2013, p. 3). Metadata variables are on a more macro level,[10]exhibiting little variability (ibid): e.g., the survey’s response rate is one single value. This is illustrated by considering three important categories of metadata specific to surveys:[11]

1. Descriptions of the *survey* include its name, an outline of study goals, the survey mode, and the interviewer training handbook.
2. Metadata on *items*, often in a codebook, encompass the names of the variables, possible values, interviewer instructions, and question wordings.
3. *Aggregated data* and *(statistical) summaries* can come from aggregating paradata (yielding, e.g., the overall response rate) or aggregating substantive data (e.g., the share of female respondents).

[10] We use ‘macro’ to denote aggregate/higher level and ‘micro’ for individual/low level (e.g., the level of the contact attempt). Just as microeconomics considers individual consumers, firms, etc., while macroeconomics studies countries as a whole (comprised of the former).

[11] See Groves et al. (2009, ch. 10.8), Mohler et al. (2012, p. 405, 411), Kreuter (2013, ch. 1.2), and Couper (2017b).

We notice two problems, particularly in relation to paradata. First, there is no agreement on where metadata begin: i.e., how much describing, summarizing, or, crucially, aggregation turns microdata into metadata. Overlaps and inconsistencies are thus inevitable: e.g., information on an item is metadata (about that item), but, in relation to the whole survey, is also sometimes treated as paradata. Second, often in quick succession 'data' are introduced to mean 'substantive data', and then 'metadata' are defined as "data about data",[12] implying that the second "data" in "data about data" solely refer to the substantive data—erroneously (see category 3 above and Kreuter, 2013, p. 3). In actuality, there are metadata on substantive data, metadata on paradata, and so on, although these distinctions are rarely made.[13]

Auxiliary Data Without a universally accepted definition, we follow Kreuter (2013, ch. 1.3): all data other than the substantive data, i.e., auxiliary data, include paradata.[14] 'Auxiliary' is to be taken literally: supplementary data that are meant to help.

Non-paradata auxiliary data are external to and (overwhelmingly) exist independent of the survey: e.g., administrative data on the same respondents, survey organization employee data on the interviewers, or Census data on area characteristics.

Enrichment with non-paradata auxiliary data can help substantive research (broader or more in-depth information on the very same respondents), reduce respondent burden (fewer questions necessary), and guide survey processes (adjusting contact protocols based on background from the sampling frame). Survey methodology research also benefits greatly: on factual questions, one can determine whether a response is correct by contrasting the survey response with the otherwise unknown true value provided by high-quality, 'gold-standard' auxiliary data on the very same individuals. One can then investigate the causes of erroneous answering and offer solutions (see Sect. 1).

Contextual Data are any information about an event's or an individual's context, particularly social, physical, environmental, temporal, or informational context. This also includes information about relevant reference groups (e.g., the family) and abstract concepts (e.g., local social norms and the legal environment).[15] 'Context' goes beyond recording these aspects in isolation, also considering how they interact.

[12] This original, historical definition (NAS, 2022, p. 95) also, in the first "data" (instead of using 'information'), hides the pronounced heterogeneity among metadata of which the three categories above provide only a glimpse.

[13] Similarly, Scheuren (2001)'s useful micro/macro paradata distinction is hardly used. Instead, the former are equated with 'paradata' and the latter are part of metadata (Couper, 2017b, p. 5).

[14] See also Smith (2011) and Couper (2017b, p. 2, 4). Others exclude metadata and/or paradata from this umbrella category.

[15] See, e.g., Sakshaug and Struminskaya (2023) and Wilkinson et al. (2017, p. 234f.).

Two Helpful Distinctions *Micro context* refers to a specific individual or event, whereas *macro context* is on a higher level (e.g., regional). *Internal context* is, e.g., someone's emotional state, while *external context* includes local laws.
Subject of the Context *Substantive context* is what is predominantly meant by 'context' in the larger social science literature:[16] context pertaining to substantive research questions. For example, for survey respondents asked about their cannabis consumption, the substantive context includes their parents' attitude and local legality. Survey scientists, however, also consider *survey context*: the context of conducting surveys and producing survey data, both in general and specific to a particular survey. Macro survey context can be, e.g., restrictions on freedom of expression in the survey's locale. Such information would come from appropriate auxiliary data and be included in metadata—another example of overlap among data types.

We wish to emphasize the following overlap: most if not all micro survey context is part of paradata.[17] Micro context can influence behavior at the interview or response level and thus is part of the processes producing a survey's data: e.g., how sensitive is a question to a particular respondent–interviewer pairing (Tourangeau & Yan, 2007, p. 860)? Much of micro context consists of such latent constructs. Thus, one has to rely on individuals' self-reports, interviewers' observations, and other proxy indicators. Further examples are provided in Sect. 3.

Figure 1 highlights some overarching results of Sect. 2.1. First, the data types do overlap. Second, the micro–macro consideration is useful but does not distinguish data types conclusively. Third, context in its various conceptions is part of all data types and of paradata in particular.

2.2 Paradata Definitions

Above, we discussed a first pitfall for grasping paradata: overlaps. A second challenge is that definitions in older, seminal works do not reflect the current understanding fully. There is also still no universally accepted definition (Couper, 2017b, p. 4). Third, paradata definitions in the literature even use different **definitional bases**. Some definitions even require two of them, source and content (e.g., West, 2011, p. 1 and McClain et al., 2019, p. 199). This is perhaps because of each base's weaknesses. For each of the three definitional bases (italicized boldface), we give example definitions and discuss some weaknesses below.
Source Paradata are "captured during the process" (Kreuter, 2013, p. 3). Sometimes the by-product or automated nature is emphasized (e.g., Couper, 2000, p. 393 and Roßmann & Gummer, 2016, p. 313).

[16] Also, often only external and macro context is considered—without even addressing these restrictions. Terms like 'environment' and 'surrounding' may unnotedly induce too narrow notions of 'context'.

[17] This was perhaps first articulated very explicitly by Kreuter (2018a, p. 193).

However, substantive data are also captured during the survey process, and some paradata variables are not but derived later (see Sect. 3). The by-product or automatic nature is missing from, e.g., interviewer observations.

Content Paradata are "describing" or "about" the process (Couper, 2000, p. 393; Nicolaas, 2011, p. 1).

Yet, some paradata are themselves not about any process directly: e.g., raw audio recordings, observed neighborhood characteristics, or whether respondent and interviewer have the same gender (as an aspect of their interaction).

Use Paradata are "used to manage and evaluate the survey process" (Couper, 2017b, p. 4f. on Groves & Heeringa, 2006).

However, sampling frame information and other auxiliary data are also used to "manage and evaluate the survey process", and substantive research, too, employs paradata. Taken literally, absolutely nothing would be paradata unless and until it has been actually used to "manage and evaluate".

Process is a common refrain among paradata definitions. We find the singular 'process' misleading: it may be the reason why some equate the whole survey process with only the field phase, the data collection, or even only the interviewing process (see Couper, 2017b, p. 4, on paradata's narrow origins). Thereby neglected *processes* include the design phase, postprocessing (editing, labeling, and coding), and two repeated processes: recruitment and the question–response process. The latter itself comprises comprehension, retrieval, judgment/estimation, and reporting processes (Groves et al., 2009, ch. 7.2). All these processes, and their complex relations, influence the survey as a whole and a specific cell's value in the released substantive data.

From authors of other chapters, we learned that some of their fields struggle with how the terms 'process data' and 'paradata' relate exactly. In our field, there is near-universal agreement that all paradata are process data.[18] The reverse question is less settled: some disagree that all process data are also paradata (e.g., Lyberg, 2011, p. 8), whereas some agree although they often equate the terms just for their paper (e.g., Kreuter et al., 2010a, p. 282 and 286). The former do not provide counterexamples. Unfortunately, neither define 'process data'.

We surmise that some processes, happening in temporal or spatial proximity to survey processes, produce process data, but not *survey* paradata:[19] e.g., internal processes such as human resources of the survey organization, or the processes of processing, analysis, and algorithmic decision-making (Enqvist 2024) on the released substantive data.

[18] We know of only two, early exceptions that did not catch on: very wide definitions that include other auxiliary data, thus going beyond 'process' (Kennickell et al., 2009, p. 1: sampling frame), and those distinguishing routine/process paradata from added/(interviewer-)observational paradata (Smith, 2011, p. 1f.).

[19] Perhaps unless repurposed (see definitional base *Use*): e.g., billing information from traveling interviewers may be used for cross-checking the so-called call records (see Sect. 3).

Our Definition is to reflect the heterogeneity of what are and what is seen as paradata. It synthesizes existing definitions. Survey paradata are data

1. that are themselves not the survey's substantive data, and
2. that would typically not exist without the particular survey, at least in the particular form available, and
3. that were automatically produced, actively collected, or constructed later on, and
4. that relate to the survey and its processes in at least one of three ways:
   a. Data produced by the survey processes, often as by-products
   b. Data describing the survey processes, including proxies for unobserved constructs and (micro-)contextual information about the survey processes
   c. Data used to manage and evaluate the survey process(es).

3 Paradata Examples

Within each category (boldface), we usually first present primary, raw paradata and then *some* 'derived variables', i.e., created from the former or other data sources. The categories are to facilitate understanding. They partly overlap.

Timing is first captured as time stamps (time and date) from which much can be derived: on which day of the week is the interview, is it a holiday, or how much time has passed since the start of the field phase, the last interview, etc. Response times are how long it takes a respondent(-interviewer paring) to complete a specific item in a particular survey (Matjašič et al., 2018); these times add up to the interview duration.

Call Records are kept about prior contact attempts for each sampled unit. Note that survey scientists call contact attempts 'calls', regardless of survey mode. Together with each call's outcomes (**disposition codes**: noncontact, rescheduled, completed, ...; reasons for refusal), they are also termed **contact history data**. **Recruitment phase data** are the web survey analogue (McClain et al., 2019, p. 200f.). Much information can be derived: e.g., a unit's current status; level-of-effort measures (Olson, 2006, p. 744f.); contact sequences (Durrant et al., 2019); and response histories in panels[20] (Kreuter & Jäckle, 2008).

Audio, Verbal, or Voice Paradata comprise recordings and features automatically extracted in real time. Derivable variables include pitch, speed, disfluencies—particularly their levels, changes, and the respondent–interviewer similarity; over-

[20] 'Panel surveys' interview the *same* respondents at multiple points in time ('waves'), e.g., annually.

speech; and whether a question was misread by the interviewer (Jans, 2010; Conrad et al., 2013; Olson & Parkhurst, 2013, ch. 3.3.5).

Location Paradata can come from, e.g., GPS (Edwards et al., 2017, ch. 12.3), other devices (Keusch et al., 2024), or IP addresses (Felderer & Blom, 2022). Interviewer travel distance and patterns or whether the respondent was on the move during the interview are examples of dynamics that can be derived.

Device Paradata mainly concern web surveys: e.g., device type (PC, smartphone, tablet), operating system, and browser settings (Callegaro et al., 2015, ch. 2.4.2.2).

Human Interface/Input Device Paradata mostly come in two forms: **Keystroke data** log each key pressed by the interviewer and the respondent. For example, sequences and how often the help/back/delete keys were pressed can be derived. **Mouse tracking** captures a computer mouse's movements and clicks, yielding timed sequences of coordinates and events. They allow the calculation of distance traveled by the mouse cursor, deviation from the direct path, velocity, and acceleration, as well as hovers over response options (Kieslich et al., 2019; Fernández-Fontelo et al., 2023). Both forms inform about navigation, idle times, and whether and when responses were changed and what the previous answer was. Analogues for smartphones and tablets have been developed (Schlosser & Höhne, 2020).

Interviewer Observations can be about, e.g., the neighborhood (signs of vandalism), dwelling (the presence of children or whether interviewer access was blocked by a gate), person, or interview (interruptions by children). **Interviewer ratings** are evaluations of, e.g., the respondent's interest, effort, or satisfaction, and the interviewer–respondent interaction (Kirchner et al., 2017; Jacobs et al., 2020).

Respondent's Ratings and Self-Ratings mirror interviewer ratings. Respondents are either explicitly prompted for their ratings or can provide information about their particular survey in an 'open comments' section at the end.

Interviewer Characteristics can be either fixed or varying. The former are sometimes seen as paradata: sociodemographics, position or experience in the organization, etc. (via employee data); or attitudes, traits, education, skills, and years working as an interviewer (via interviewers answering a separate questionnaire). Varying characteristics need to be calculated: the number of prior calls or completed interviews on the same day or during the field phase overall, time since the last interview, etc.

Few fixed **respondent characteristics** are widely considered paradata, except for some attitudes (about being interviewed, scientific surveys generally, this survey's topic) and prior survey experience (Matthijsse et al., 2015; Schwarz et al., 2022, ch. 2). Varying respondent characteristics are discussed under Interactions below.

Survey/Interview Characteristics such as incentives, recruitment strategies, and offered mode can vary: across units, time, or in multicountry efforts.[21]
Item Characteristics inevitably differ between items: e.g., length, response options, and topic. But even a particular item may be different across respondents: e.g., because of adaptations based on the participant's prior responses.

Interactions are often only accessible via proxies or subjective judgments. For each *respondent–interviewer* pairing, ratings can be captured, the difference in age, attitudes, or language can be derived, or specific aspects of this complex social interaction can be addressed (Bradburn, 2016). For the *respondent-survey* interaction, reasons for participating (Schwarz et al., 2022, ch. 2.2) or whether it was conducted in the respondent's native tongue are examples. The *respondent-item* interaction, too, contains subjective aspects, e.g., item sensitivity or trouble understanding, and objective aspects, e.g., the number of all or of similar questions answered before.

Micro Survey Context partly overlaps with observations, ratings, and interactions. Adding to Sect. 2.1, we highlight some further latent constructs (italicized) and respective proxies. Perceived level of *privacy* (Yan, 2021, p. 120): Was the respondent at home or in a public space? Who else was present: a boss, spouse, or children? Trust and interviewer–respondent *rapport* (Sun et al., 2021): Was it always the same interviewer, in panels or in continued recruitment attempts (Kühne, 2018)? *Engagement* and *effort*: Did the respondent multitask? Did they look up information in documents?

Design Phase[22] paradata include changes made after pretesting a survey (see Sect. 5). Online comments are a simple tool for volunteering information: e.g., respondents about comprehensibility, offensiveness, or other issues with a question, or experts about design flaws (Callegaro et al., 2015, p. 105, 109).
Editing/Coding[22] paradata about each cell of the substantive data can be whether the value came from the respondent or a labeler (Sana & Weinreb, 2008). More detailed information for the latter case includes the labeler id or the rate of agreement among multiple labelers looking at the same cell.

Miscellaneous Video recordings, eye-tracking measures, and brain activity data are rare because of equipment requirements (Callegaro et al., 2015, p. 108f.).

Some surveys inquire about providing biosamples, willingness to be contacted again, or allowing record linkage to other data. The respondent's consent decision or reasons for refusal (Sakshaug, 2013) may be indicative of respondent behavior.

The status or relation of who provided information can be relative to the sample unit (targeted person versus family member) or to the information (information

[21] If fixed they are considered metadata.

[22] These are examples of paradata not accruing in the field phase. They are rarely released.

provided about oneself or about someone else). In establishment surveys, the respondent's position within the company might influence their response.

Sensors, wearables, and apps have received attention recently (Keusch et al., 2024), but only a part of these data are paradata.

4 Collecting, Structuring, and Designing Paradata

Below, we consider some differences in how paradata types are collected and structured. This is not just inherent heterogeneity: many paradata types can also be actively designed. Thus, paradata are not necessarily 'found' or 'organic data' (Groves, 2011; Japec et al., 2015, p. 843) over which researchers have no discretion.

Resolution *Device-recorded paradata* are usually constrained only theoretically, as technical resolution limits are beyond sufficient.[23] Resolution for response time should be at least 100, better 10 or 1 millisecond(s) (Mayerl, 2013, p. 3): if differences across individuals ('signal') are on the order of seconds, then measuring only up to the second ('noise') degrades too much information (signal-to-noise ratio)—needlessly.

Human-recorded paradata should heed survey methodology's advice on how to construct items and response scales (e.g., Bradburn et al., 2004; Groves et al., 2009, ch. 7).[24] Here, higher resolution can be detrimental: e.g., contact history systems with very many disposition codes (AAPOR, 2016, p. 71ff.) may produce minuscule counts for some outcomes and errors (e.g., by inexperienced interviewers). If one anticipates combining categories later on, then the design should facilitate this.

Granularity Ratings may reflect the whole interview, segments, or use, e.g., 'increasing/steady/declining' to capture dynamics. Response time can be at the level of questionnaire sections, the page (web mode), item, or even finer (see Components). Item-level analysis is impossible when measurements are only at the page level.

In web surveys, server-side (Callegaro, 2013, p. 262) measurements can always be implemented but are restricted to the page level and (differential) transmission, and loading times are unwanted components—in contrast to client-side measurements (i.e., collected on respondents' devices) whose collectibility, however, depends on user consent and device (Callegaro et al., 2015, ch. 5.3.4.1).

Components: Splitting/Combining The idealized question–answer sequence comprises an interviewer reading aloud, the respondent's cognitive processing,

[23] Data volume may be a constraint for transmitting and storing audio or video recordings.

[24] For example, for ratings: a 5-point scale, each point labeled, ordered from low to high, and equidistant.

their answering, and the entering of the response. Thus, item-level response time can be split into four components, yielding more nuanced information.

Aggregation This often refers to the level at which a variable operates (varies) within the complex, hierarchical survey structure: item, respondent, interviewer, call, or interview. Appropriate aggregation reduces informational overload (see Sect. 6) and enables targeted applications: e.g., interviewer level monitoring and item level design evaluation (see Sect. 5).

Ex post, one can usually decrease granularity, combine components,[25] and aggregate. The reverse direction is typically impossible: information not captured is lost forever.

Degree of Automation Some paradata are always collected automatically (e.g., keystrokes) and some are never (e.g., interviewer observations). For others, such as response times, there are several options:

1. Manual, 'active' time stamps: The interviewer presses a button after having read the question aloud completely and again when the participant starts responding.
2. General automation: The timer always starts when the question appears on-screen and stops when the response is confirmed.
3. Specific automation: A voice-activated system recognizes when speaking starts and stops.

Each approach has its advantages. Interviewers (1) know best when they finish and can also ignore nonanswers (e.g., thinking out loud or asking for clarification); they can also record whether a measurement was valid. Meanwhile, automation frees up the interviewer. General automation (2) is unsusceptible to inadvertent button-pressing and nonanswering but combines all interviewer and respondent components into one measurement. Specific automation (3) can separate them but is hampered by overspeech, low volume, nonanswering, and needs specialized software. Combined, semi-automated versions try to reap the benefits of each approach.

Raw Paradata are sometimes not fit for the intended use.

Preprocessing turns raw mouse-tracking data—a continuous stream of events, coded in computer language—into comprehensible, usable information (Olson & Parkhurst, 2013, ch. 3.3.3, 3.5.1). Specialized software exists (Wulff et al., 2021; Henninger et al., 2022b), unlike for processing tasks needing human labelers such as rating recorded interactions. Other paradata may need trivial (response time = stoptime − starttime) or no work (interviewer ratings).

[25] Problematic are components that are overlapping or nonsequential. For instance, summing response times over the seemingly additive four question–answer process components produces an overestimate when the participant starts responding before the interviewer finishes the question. Clever programming can sometimes solve such problems.

Adjustment shall denote the correction for unwanted properties. Raw response times are influenced by characteristics of the respondent (e.g., their general baseline speed; Mayerl et al., 2005), interviewer, item, device, and so on (Couper & Peterson, 2017; Sturgis et al., 2021, ch. 1). Response times become comparable only after accounting for such influences; otherwise, it would remain unclear if someone was speeding on an item or is just generally fast. This is usually done by statistical regression (Couper & Kreuter, 2013), which is only possible after data collection, i.e., not in real time. Further examples benefiting from adjusting for respondent idiosyncrasies include verbal and mouse paradata.

Capturing Paradata can take on different, potentially complementary forms. *Location* is, mostly, about server-side and client-side in web surveys (see Granularity). The *origin* of ratings can be from interviewers, respondents, or labelers, of observations from interviewers, recruiters, or listers, and of response times from human interviewers or several devices.

Availability for all relevant units and at a point in time limits applications (see Sect. 5).
When While interviewers continually observe the process, their evaluations are only recorded after the interview's conclusion. Other paradata are available in (near) real time—even some derived variables: e.g., idle times and response editing, from keystrokes or mouse tracking. However, a need for nonautomated or time-consuming preprocessing or adjustments impedes real-time interventions (Mittereder, 2019, p. 153). On a different note, early in the field phase the paucity of paradata limits applications (West et al., 2023).
On whom There are more data for completed cases (e.g., ratings) than for breakoffs (anything collected until termination), contacts (outcome codes; some interviewer observations), and noncontacts (GPS; neighborhood observations) (Sakshaug & Kreuter, 2011).

5 Applications

Paradata are employed for the various survey methodological challenges (see Sect. 1): errors of representation, errors of measurement, and missing data, i.e., missing units/rows ('unit nonresponse') and missing responses/cell entries ('item nonresponse'). The first goal is to recognize errors and the underlying mechanisms. Then, the design of future surveys can be improved. Also, for given survey data, statistical methods such as imputation and weighting can be used to derive unbiased results even from deficient data. The second goal is to monitor data quality and to predict problems: this is the basis for interventions.

Any statistical modeling of behavior is constrained by what paradata are available for all relevant units. Prediction, in addition, is restricted by information available at prediction time (see 4).[26]

Unit Nonresponse is likely the most studied application. Other than the (very limited) sampling frame information, paradata may be all that is available for respondents and nonrespondents alike (Sinibaldi et al., 2014):[27] e.g., observations about the neighborhood, dwelling, or individual(s), call histories, and interviewer characteristics such as their voice (Kreuter & Casas-Cordero, 2010, p. 3; Olson, 2013; Charoenruk & Olson, 2018).[28]
Avoidance of nonresponse bias builds on increasing the recruitment effort, monetary incentives (Jackson et al., 2020), or the many adaptive survey design strategies (see below) on cases predicted to be difficult or important for sample representativity.
Adjustment for nonresponse in the eventual data analysis of the substantive data usually involves some form of weighting based on the response propensity P. To repair nonresponse bias, a paradata variable employed in estimating P must be strongly correlated with both P and the survey variable of interest (Kreuter et al., 2010b). However, rarely a single available paradata variable exhibits enough correlation with both; using multiple variables may help (Kreuter & Olson, 2011).
Panel dropout of participants between waves of a panel survey can be studied with prior waves' paradata: comments (McLauchlan & Schonlau, 2016); response behavior and speed (Roßmann & Gummer, 2016); interviewer observations (Plewis et al., 2017); and habitual late responding (Minderop & Weiß, 2023). *Breakoffs* are more frequent in web mode, on mobile (versus PC) and nonpreferred devices, and preceding response behavior such as speeding and instability is predictive (Mittereder, 2019; Couper et al., 2017; Chen et al., 2022; Mittereder & West, 2022).

Coverage Error can be addressed in two ways (Eckman, 2013). First, two sources can be compared. For example, sampling frame versus a 'lister' walking a neighborhood to collect addresses and contact information: flagging, additions, and deletions of units. How much self-reports, sampling frame information, or interviewer observations match on survey inclusion criteria is an indicator of their accuracy. Second, whether the particular circumstances affected sampling frame creation is of interest: e.g., duration, weather, time, location data, lister's or interviewer's discretion, and edits.

[26] This includes the subtle 'data/target leakage' (Ghani & Schierholz, 2020, ch. 7.8.1).

[27] Ditto for what is available before contacts are attempted.

[28] There is more information for refusals and dropped contacts (Sakshaug & Kreuter, 2011).

Device paradata can inform about the error that would be introduced when survey participation required apps available only for some smartphone models (Couper et al., 2017, ch. 7.2). Similarly, to increase representativeness, some online panels have offered free devices and Internet to those lacking (Blom et al., 2017). Paradata on who was such an 'offliner' allow studying such programs' success regarding participation and improving substantive results (Cornesse & Schaurer, 2021; Eckman, 2016).

Errors of Measurement (Yan & Olson, 2013) and **Item Nonresponse** are often studied jointly as both they concern (error-prone and missing, respectively) cell entries in the substantive data.

Paradata measure or proxy behaviors, context, and mechanisms that influence these data quality aspects: device (Lugtig & Toepoel, 2016); multitasking (Sendelbah et al., 2016; Höhne et al., 2020b); regional context (Purdam et al., 2020); rapport (Sun et al., 2021); consistency of related answers (Revilla & Ochoa, 2015); uncertainty, slow or fast responding, changing of responses, soliciting help, interviewers misreading (Yan & Olson, 2013); ratings (Holbrook et al., 2014; Olson & Parkhurst, 2013, ch. 3.3.6); verbal paradata (Jans, 2010); reasons for participating such as incentives (Matthijsse et al., 2015; Schwarz et al., 2022); and interviewer characteristics influencing the sensitivity of a specific question (Peytchev, 2012). Respondent self-reports provide additional information to that already contained in other paradata (Revilla & Ochoa, 2015; Höhne et al., 2020a).

Adaptive Survey Design (ASD) and **Responsive Survey Design** (RSD)[29] are popular, mostly to lower costs and increase data quality (e.g., Wagner et al., 2012). One perspective is that the harder a unit is to recruit, the more similar it presumably is to nonrespondents (Olson, 2013, p. 155). Thus, when data collection stabilizes, i.e., primary substantive variables do not change anymore with increased contact attempts, one may move to another RSD phase, tweak protocols, or stop data collection. *Real-time interventions* in ASD can be appropriate pop-up messages to prevent breakoffs (propensity predicted with paradata: Mittereder, 2019, ch. 6) or slow down speeders (Conrad et al., 2017). Offering clarifications in self-administered surveys based on age-adjusted idle time can improve response accuracy and satisfaction (Conrad et al., 2007). Allowing the interviewer to ask only the most important questions when they predict a high risk for unit nonresponse or breakoff is more drastic (Lynn, 2003).

[29] See Schouten et al. (2017, ch. 2.2). *ASD* (Wagner, 2008) aims to tailor survey design to individuals/groups: e.g., case prioritization (with paradata-predicted propensities: Wagner et al., 2012), call timing (time successful in previous wave: Kreuter & Müller, 2015), contact strategies, mode, and interviewer type (Tourangeau, 2021, ch. 2). In contrast, *RSD* (Groves & Heeringa, 2006) tries different designs in prespecified early phases of the field phase to arrive at an optimal final design.

Monitoring and Evaluation guide the complex survey processes in real time (Couper, 2017b, p. 10). Dashboards (Mohadjer & Edwards, 2018, p. 263ff.) visualize information for survey managers: in particular, 'key performance indicators' of costs, data quality,[30] and interviewer performance[31] (Meitinger et al., 2020).

Performance can improve with feedback to interviewers when data sources conflict (GPS vs. call history about locations: Edwards et al., 2017, ch. 12.3; Wagner et al., 2017, p. 221) or when paradata indicate deviations from protocols (Edwards et al., 2020). Recordings can be reviewed for quality control. Interview durations may inform about deviant behavior (fabricated interviews: Schwanhäuser et al., 2022).

Evaluation of Survey Design in the evaluation phase (Maitland & Presser, 2018), in pretesting (Couper, 2000; Stern, 2008), by experts (comments about items: Callegaro et al., 2015, p. 105, 109), and during the field phase uses paradata to indicate problems: slow responding, rates of item nonresponse and changed responses, going back to earlier related items, interviewer evaluations, and labeler-coded behavior.

Costs not being available in real time or in full detail hampers survey administration. Then, estimating cost parameters from call histories may help (Wagner, 2019).

Substantive Research has used survey paradata, too, but is beyond the focus of this survey methodological chapter. For example, interviewer observations can supplement the substantive data when missing or for quality control: e.g., the presence of wheelchair ramps and cigarette butts in health surveys (West, 2018a, p. 212).

Response times or 'latencies' have long been used to study cognitive processes such as the degree of elaboration (deliberative-controlled or automatic-spontaneous processing), abilities, strength of attitudes, and mental availability of information (Johnson, 2004; Mayerl, 2013; Kyllonen & Zu, 2016; De Boeck & Jeon, 2019).

6 Challenges and Some Solutions

Paradata Quality is understudied in general (West & Sinibaldi, 2013). Interviewer-produced paradata have received relatively more attention (Olson, 2013, p. 159). Automation (see 4) and objective paradata (West & Sinibaldi, 2013, ch. 14.2.3) do not guarantee high(er) quality.

Errors, including a lack of internal validity or reliability, in interviewer observations have been noted at rates between $<10\%$ and 92% (West, 2011, p. 4; West,

[30] 'Representativeness indicators', using contact history paradata (Schouten et al., 2012).

[31] Accounting for each interviewer's case difficulties with call histories (West & Groves, 2013).

2013b). Context (e.g., seasonality, cooperation, sensitivity) and characteristics of the respondent, household, area, and interviewer can influence interviewer observations (West & Li, 2019; West & Blom, 2017, ch. 4.8). Unfortunately, performance may actually decline during the field phase (West & Sinibaldi, 2013, p. 351). Also, inter*labeler* reliability can be challenging (verbal paradata: Jans, 2010, ch. 2.2).

Missing Values can be frequent in, e.g., interviewers' neighborhood observations (Olson, 2013, p. 146) and call records (Wagner et al., 2017, ch. 5.3). Reasons include ambiguous guidelines or cases (Biemer et al., 2013), forgetting when recording later,[32] and hesitancy to record sensitive information. Also, using multiple devices can hinder completeness (Höhne et al., 2020a, p. 994).

Solutions for improving general survey quality are also informative for survey paradata quality. For instance, operationalizations of interviewer observations should heed survey methodology's general lessons better (see 4 and Kreuter, 2018b, p. 534). Systems must facilitate easy, timely entry (ibid), with errors easy to correct or flag: e.g., interviewers can rate each response time measurement as valid, respondent error, or interviewer error (Mayerl, 2013, step 2.2). Automatic consistency and completeness checks (West & Sinibaldi, 2013, p. 352f.) can compare across data sources (for location: from GPS vs. from call records) or to normal values (unusual response times: West, 2013a, p. 352f.): after all, interviewers can prevent or correct problems best in real time. Frequently recommended are standardized, high-quality, survey-specific training and, periodically or when needed, retraining of interviewers, reminders, checklists, instructions, and the like (e.g., Kreuter, 2018b, p. 534).

Informed Consent about paradata collection is an ongoing debate (Connors et al., 2019, p. 187f.). Should respondents be informed—and how?[33] Do they need to consent (Kunz et al., 2020a, p. 397f.)—at all,[34] as part of one overall agreement to participate in the survey, or in a separate paradata consent question?[35] Nonconsent may reduce participation (Couper & Singer, 2013) and bias samples (Felderer & Blom, 2022, p. 878), although much less so when following the emerging best practices (Kunz et al., 2020b). We would like to caution that in the long run a lack of transparency could backfire and reduce trust and participation rates.

[32] Interviewers may not record information right away, even when they are supposed to, due to, e.g., safety concerns or a lack of time (West & Sinibaldi, 2013, p. 346f.; Kreuter, 2018b, p. 533).

[33] Couper and Singer (2013) favor informing about usage rather than about paradata themselves. The wealth of applications (see Sect. 5) and lack of control about how released paradata are used make this questionable. Informing about both, paradata types and paradata uses, may be necessary.

[34] Requiring consent is reasonable but hard in practice: e.g., nonrespondents can hardly consent.

[35] Paradata, being largely invisible, are different from the substantive survey questions. The respondent gets to know each of these and can, for each item, choose not to respond or to discontinue.

Confidentiality concerns are highest in, e.g., address details, interviewer observations, open-text answers, and recordings (Nicolaas, 2011, p. 15). Selective anonymization is hard for unstructured paradata (Kreuter, 2018b, p. 535). The general approaches for sensitive data (see Shlomo, 2018 and Bender et al., 2020) can be solutions for paradata, too. Also, paradata may be used in real time, never leaving a respondent's device (Henninger et al., 2022a, p. 16). Also, *perceived* privacy (Nicolaas, 2011, p. 15), the actual driver of consent and behavior, must reflect reality.

Availability of Paradata is hampered by organizations guarding internal best practices, resources needed for preparation, warehousing, and documentation (Nicolaas, 2011, p. 16 and 14; Olson, 2013, p. 162), and confidentiality questions (Kreuter, 2018b, p. 535). (Micro) Paradata are released more frequently nowadays but often only contain some of the paradata variables or only the completed interviews. Research about paradata may also stay internal for similar reasons or because improvements are deemed small (Wagner, 2013b, p. 166).

Standardization may be helped indirectly by the dominance of a few software solutions (web: McClain et al., 2019, p. 201f.).[36] Yet, in contrast to metadata there are almost no universal paradata standards (Vardigan et al., 2016, p. 445; Couper, 2017b, p. 7). Even within an organization there may be heterogeneity on how to record information: e.g., among interviewers with different experiences at prior employers or between survey methodologists and interviewers. Concrete, clear standards are key. Yet, standardization must leave room for tailoring paradata (Kreuter, 2018b, p. 534): e.g., to specific contexts and needs (see West & Sinibaldi, 2013, p. 347 and 5 on nonresponse adjustment variables having to fit the specific application).

Overwhelming users is a common worry about paradata (Couper, 1998, p. 45; Kreuter et al., 2010a, ch. 5). This is in part, but not only, about volume.

The *informational content per observation* is, however, only high for some variables: e.g., an interviewer's exhaustive free-text call notes may be useful to themselves but overwhelm other follow-up interviewers or managers (West & Sinibaldi, 2013, p. 347). Standardizing and structuring the minimum informational content while making additional notes optional is an easy fix. *Many paradata variables* are or might be available. Beginning with those that one knows will be used and for which one has applications is a great starting point (West, 2018a, p. 213). Some paradata variables have *many data points*: e.g., every single mouse coordinate.

[36] However, the reliance on a few, underfunded or volunteer 'research software engineers' is very worrisome, especially as these systems must coevolve with technology and society.

Instead of appraising every single, microlevel value, information is aggregated to the appropriate level, reduced in dimension (e.g., by clustering) or to special cases (e.g., outliers), or fed into statistical methods.

Handling Paradata can seem daunting at first. Yet, the *separate files* for call records, interviewer characteristics, and item-level paradata can be merged. *Levels* may be changed by aggregation, or, in files, by 'reshaping' between long and wide data formats. All this is facilitated by software and need not be done manually.

The *structure* of many paradata variables can be nontrivial. Where detailed statistical analysis of paradata is needed, hierarchical, complex structures are addressed with multilevel modeling.[37] Call records are an example of unbalanced data: zero, one, or more observations per unit. Yet, this is only sometimes actually problematic. Then, simple aggregation is often sufficient: e.g., counts per unit. There are also less crude methods that can target patterns as a whole, e.g., in call histories and mouse movement trajectories (Durrant et al., 2019; Fernández-Fontelo et al., 2023).

Heterogeneity abounds across cases. Some accrue more information (completed interviews) or more observations (repeated calls). One variable may capture different concepts (Olson, 2013, p. 159): attempts made (nonrespondents) or calls needed for success (respondents). In surveys with multiple modes (e.g., ASD and RSD), some variables are not available in each or not directly comparable (Kreuter, 2018a, p. 195).

Information Is Lacking on many processes (respondents' and interviewers' true motivation, states, and behavior) or because of too few cases (e.g., breakoffs and fabricated interviews). Unsupervised learning (James et al., 2021, ch. 12) may help: e.g., clustering for finding deviant interviewer behavior (Schwanhäuser et al., 2022).

Misalignment of Incentives between, e.g., interviewers and survey designers or researchers, can be problematic. Yet, studies of, e.g., prevalence and reasons for interviewers ignoring recommendations are rare (call timing: Wagner, 2013a, Experiment 5; travel routes: Tourangeau, 2021, p. 17f.). Remuneration schemes ignoring the time needed to record paradata[38] clash with expectations for high-quality paradata.[39] With (perhaps diffuse) monitoring, interviewers may feel the need to demonstrate performance (West & Sinibaldi, 2013, p. 343, 347). Transparency is a partial solution (West & Groves, 2013, p. 373): letting the interviewer

[37] See Couper and Kreuter (2013) on response times. Multiple observations belonging to the same unit or interviews by the same interviewer are correlated and not isolated, independent data points.

[38] Time requirements can be sizable: e.g., 15–20 minutes per interview (West, 2018b, p. 541).

[39] Conversely, West and Sinibaldi (2013, p. 344) did not find that rewarding each contact attempt induced interviewers to overreport calls.

know why they get relatively more difficult cases and that good paradata help fair evaluation.

Overall, one may need to convince the interviewers of the value of quality paradata, in general and to themselves, via improved case assignments and improved recommendations (West & Sinibaldi, 2013, p. 348; West, 2018a, p. 212). The same is true for survey managers, listers, recruiters, and other actors on the ground or in decision-making positions (Olson, 2013, p. 161).

7 Discussion

(Un)intended Consequences of making paradata and paradata collection explicit need further study. Changed behavior among "watched" respondents is plausible but has not been found yet (Kunz et al., 2020a, p. 402) except for participation (Henninger et al., 2022a, p. 5f., 9). When recorded, interviewers produce fewer suspiciously short durations (Olbrich et al., 2022). On a different note, making interviewers predict respondent behavior could yield self-fulfilling prophecies (Eckman, 2017, ch. 3).

Perspectives on paradata are many and varied. This is true across disciplines, as this volume shows, but also within our field. Most research has started from either the available paradata or established knowledge about surveys. Those on the ground—labelers, field staff, interviewers (Jans, 2010, ch. 2.2; West & Sinibaldi, 2013, ch. 14.2.2.1; West & Trappmann, 2019)— have hitherto untapped knowledge about processes, their own strategies, and working with researchers' paradata instruments.

Ethics and critical reflection of potential harm from paradata collection and applications are paramount (AAPOR, 2021). Survey methodology is shaped by mostly benign surveys. In the West or elsewhere, respondents and interviewers from some locales, contexts, or specific groups are rightfully afraid of negative consequences from honest answering or mere participation. Yet, many of the ethical and legal struggles (see also Sect. 6) are not unique to paradata (Conrad et al., 2021, p. 254).

Costs and Trade-Offs relate questions of data quality to each other and to the real world. Paradata may be by-products—they are not why surveys are conducted—but they are not cost-free: Systems need development and maintenance; recording information (interviewers), monitoring quality (managers), and training (both) take time and effort; paradata must be preprocessed and documented before being released. That paradata are high-quality is not a given, either (see Sect. 6).

Our field does not have a common framework for all survey costs and few empirical studies on utility per dollar. Trade-offs are recognized but hard to quantify. Resources spent on paradata basics (e.g., infrastructure) cannot be spent to improve one survey's substantive data (quantity or quality) but can benefit many future surveys.

We have discussed many examples and challenges to provide a broad overview, but one important message should not get lost: some paradata types are easy to capture and contain much information relative to the resources that must be invested.

Take a Paradata Perspective When Helpful Whether everyone agrees that something is paradata or whether they would, had it been created differently, will not diminish its usefulness. Paradata are not an end unto themselves, but "additional [. . .] tools" to help in practice (Couper, 2017b, p. 11), not meant to replace other tools or perspectives. *Use* may not seem the most important definitional base for paradata (see Sect. 2.2), but, after all, applications are why we capture paradata.

Acknowledgments The authors would like to thank the editors, two anonymous referees, the other authors, and members of the FK2RG research group for their comments. The authors would also like to thank the editors for an extraordinarily smooth, accommodating, and stimulating process.

References

AAPOR (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (9th ed.). The American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2022/11/Standard-Definitions20169theditionfinal.pdf

AAPOR (2021). *AAPOR Code of Professional Ethics and Practices*. The American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2022/12/AAPOR-2020-Code_FINAL_APPROVED.pdf. Revised April 2021.

Bender, S., Jarmin, R. S., Kreuter, F., & Lane, J. (2020). Privacy and confidentiality. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science* (2nd ed., Chap. 12). CRC Press. https://textbook.coleridgeinitiative.org.

Biemer, P. P., Chen, P., & Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 147–168.

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., & Bossert, D. (2017). Does the recruitment of offline households increase the sample representativeness of probability-based online panels? Evidence from the German Internet Panel. *Social Science Computer Review, 35*(4), 498–520.

Bradburn, N. M. (2016). Surveys as social interactions. *Journal of Survey Statistics and Methodology, 4*(1), 94–109.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design*. Jossey-Bass, Wiley.

Callegaro, M. (2013). Paradata in web surveys. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information*. Wiley.

Callegaro, M., Manfreda, K. L., & Vehovar, V. (2015). *Web survey methodology*. Sage

Charoenruk, N., & Olson, K. (2018). Do listeners perceive interviewers? Attributes from their voices and do perceptions differ by question type? *Field Methods, 30*(4), 312–328.

Chen, Z., Cernat, A., & Shlomo, N. (2022). Predicting web survey breakoffs using machine learning models. *Social Science Computer Review, 41*, 573–591.

Connors, E. C., Krupnikov, Y., & Ryan, J. B. (2019). How transparency affects survey responses. *Public Opinion Quarterly, 83*(S1), 185–209.

Conrad, F. G., Broome, J. S., Benkí, J. R., Kreuter, F., Groves, R. M., Vannette, D., & McClain, C. (2013). Interviewer speech and the success of survey invitations. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 191–210.

Conrad, F. G., Keusch, F., & Schober, M. F. (2021). New data in social and behavorial research. *Public Opinion Quarterly, 85*(S1), 253–263. Introduction to Special Issue: New Data in Social and Behavioral Research.

Conrad, F. G., Schober, M. F., & Coiner, T. (2007). Bringing features of human dialogue to web surveys. *Applied Cognitive Psychology, 21*(2), 165–187.

Conrad, F. G., Tourangeau, R., Couper, M. P., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods, 11*(1), 45–61.

Cornesse, C., & Schaurer, I. (2021). The long-term impact of different offline population inclusion strategies in probability-based online panels: Evidence from the german internet panel and the GESIS panel. *Social Science Computer Review, 39*(4), 687–704.

Couper, M., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 271–286.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. In *Proceedings of the Survey Research Methods Section of the American Statistical Association, American Statistical Association* (pp. 41–49). Joint Statistical Meetings of the American Statistical Association.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review, 18*(4), 384–396.

Couper, M. P. (2017a). New developments in survey data collection. *Annual Review of Sociology, 43*, 121–145.

Couper, M. P. (2017b). Birth and diffusion of the concept of paradata. *Advances in Social Research, 18*. https://www.jasr.or.jp/english/JASR_Birth%20and%20Diffusion%20of%20the%20Concept%20of%20Paradata.pdf. English manuscript by Mick P. Couper, page numbers refer to pdf file.

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys. In P. P. Biemer, E. D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 133–154). Wiley.

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review, 35*(3), 357–377.

Couper, M. P., & Singer, E. (2013). Informed consent for web paradata use. *Survey Research Methods, 7*(1), 57–67.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 1–11.

Durrant, G. B., Smith, P. W., & Maslovskaya, O. (2019). Investigating call record data using sequence analysis to inform adaptive survey designs. *International Journal of Social Research Methodology, 22*(1), 37–54.

Eckman, S. (2013). Paradata for coverage research. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 97–116). Wiley.

Eckman, S. (2016). Does the inclusion of non-internet households in a web panel reduce coverage bias? *Social Science Computer Review, 34*(1), 41–58.

Eckman, S. (2017). Interviewers' expectations of response propensity can introduce nonresponse bias in survey data. *Statistical Journal of the IAOS, 33*(1), 231–234.

Edwards, B., Maitland, A., & Connor, S. (2017). Measurement error in survey operations management: Detection, quantification, visualization, and reduction. In P. P. Biemer, E. D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 253–277). Wiley.

Edwards, B., Sun, H., & Hubbard, R. (2020). Behavior change techniques for reducing interviewer contributions to total survey error. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (pp. 77–90). CRC Press.

Enqvist, L. (2024). Paradata as a tool for legal analysis—Utilising data on data related processes. In I. Huvila, L. Andersson, & O. Sköld (Eds.), *Perspectives on paradata: Research and practice of documenting process knowledge*. Springer.

Felderer, B., & Blom, A. G. (2022). Acceptance of the automated online collection of geographical information. *Sociological Methods & Research, 51*(2), 866–886.

Fernández-Fontelo, A., Kieslich, P. J., Henninger, F., Kreuter, F., & Greven, S. (2023). Predicting question difficulty in web surveys: A machine learning approach based on mouse movement features. *Social Science Computer Review, 41*(1), 141–162.

Ghani, R., & Schierholz, M. (2020). Machine learning. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science* (Chap. 7, 2nd ed.). CRC Press. https://textbook.coleridgeinitiative.org

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861–871.

Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Wiley.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(3):, 439–457.

Henninger, F., Kieslich, P. J., Fernández-Fontelo, A., Greven, S., & Kreuter, F. (2022a). Privacy attitudes toward mouse-tracking paradata collection. Preprint, *SocArXiv*. https://osf.io/preprints/socarxiv/6weqx/. Version from March 15, 2022.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022b). lab.js: A free, open, online study builder. *Behavior Research Methods*. Preprint at https://doi.org/10.5281/zenodo.597045

Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., & Lyberg, L. E. (2021). *Big data meets survey science: A collection of innovative methods*. Wiley.

Höhne, J. K., Cornesse, C., Schlosser, S., Couper, M. P., & Blom, A. G. (2020a). Looking up answers to political knowledge questions in web surveys. *Public Opinion Quarterly, 84*(4), 986–999.

Höhne, J. K., Schlosser, S., Couper, M. P., & Blom, A. G. (2020b). Switching away: Exploring on-device media multitasking in web surveys. *Computers in Human Behavior, 111*, 106417.

Holbrook, A. L., Anand, S., Johnson, T. P., Cho, Y. I., Shavitt, S., Chávez, N., & Weiner, S. (2014). Response heaping in interviewer-administered surveys: Is it really a form of satisficing? *Public Opinion Quarterly, 78*(3), 591–633.

Jackson, M. T., McPhee, C. B., & Lavrakas, P. J. (2020). Using response propensity modeling to allocate noncontingent incentives in an address-based sample: Evidence from a national experiment. *Journal of Survey Statistics and Methodology, 8*(2), 385–411.

Jacobs, L., Loosveldt, G., & Beullens, K. (2020). Do interviewer assessments of respondents' performance accurately reflect response behavior? *Field Methods, 32*(2), 193–212.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd Ed.). Springer. https://www.statlearning.com. First Printing August 04, 2021. Pdf Accessed August 30, 2021.

Jans, M. E. (2010). Verbal Paradata and Survey Error: Respondent Speech, Voice, and Question-Answering Behavior can Predict Income Item Nonresponse. PhD Thesis, University of Michigan, Ann Arbor, MI. https://isr.umich.edu/wp-content/uploads/2017/09/jans-dissertation.pdf

Japec, L., Kreuter, F., Berg, M., Biemer, P. P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly, 79*(4), 839–880.

Johnson, M. (2004). Timepieces: Components of survey question response latencies. *Political Psychology, 25*(5), 679–702.

Kennickell, A. B., Mulrow, E., & Scheuren, F. (2009). Paradata or Process Modeling for Inference, 2009. In *Presented at the Modernization of Statistics Production Conference, Stockholm, Sweden, 2009/11/02-04*.

Keusch, F., Struminskaya, B., Eckman, S., & Guyer, H. M. (2024). *Data Collection with Wearables, Apps, and Sensors*. CRC Press. In preparation.

Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J. M. B., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: A practical guide to implementation and analysis. In M. Schulte-Mecklenbeck, A. Kühberger, & J. G. Johnson (Eds.), *A handbook of process tracing methods* (2nd ed., pp. 111–130). Routledge. https://doi.org/10.31234/osf.io/zuvqa

Kirchner, A., Olson, K., & Smyth, J. D. (2017). Do interviewer postsurvey evaluations of respondents' engagement measure who respondents are or what they do? A behavior coding study. *Public Opinion Quarterly, 81*(4), 817–846.

Kreuter, F. (2013). Improving surveys with paradata: Introduction. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 1–9). Wiley.

Kreuter, F. (2018a). Getting the most out of paradata. In D. L. Vannette & J. A. Krosnick (Eds.), *The palgrave handbook of survey research* (pp. 193–198). Palgrave Macmillan/Springer.

Kreuter, F. (2018b). Paradata. In D. L. Vannette & J. A. Krosnick (Eds.), *The palgrave handbook of survey research* (pp. 529–535). Palgrave Macmillan/Springer.

Kreuter, F., & Casas-Cordero, C. (2010). Paradata. *RatSWD Working Papers series Working Paper No. 136, German Data Forum (RatSWD)*. https://www.konsortswd.de/wp-content/uploads/RatSWD_WP_136.pdf. Accessed Jun 24, 2022.

Kreuter, F., Couper, M. P., & Lyberg, L. (2010a). The use of paradata to monitor and manage survey data collection. In *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 282–296). Joint Statistical Meetings of the American Statistical Association.

Kreuter, F., & Jäckle, A. (2008). Are Contact Protocol Data Informative for Potential Nonresponse and Nonresponse Bias in Panel Studies? A Case Study from the Northern Ireland Subset of the British Household Panel Survey. *Paper Presented at the Panel Survey Methods Workshop, University of Essex, Colchester, UK, 2008*.

Kreuter, F., & Müller, G. (2015). A note on improving process efficiency in panel surveys with paradata. *Field Methods, 27*(1), 55–65.

Kreuter, F., & Olson, K. (2011). Multiple auxiliary variables in nonresponse adjustment. *Sociological Methods & Research, 40*(2), 311–332.

Kreuter, F., Olson, K., Wagner, J. R., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., & Raghunathan, T. E. (2010b). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*(2), 389–407.

Kühne, S. (2018). From strangers to acquaintances? Interviewer continuity and socially desirable responses in panel surveys. *Survey Research Methods, 12*(2), 121–146.

Kunz, T., Landesvatter, C., & Gummer, T. (2020a). Informed consent for paradata use in web surveys. *International Journal of Market Research, 62*(4), 396–408.

Kunz, T. C., Beuthner, C., Hadler, P., Roßmann, J., & Schaurer, I. (2020b). Informing about web paradata collection and use. *GESIS Survey Guidelines, GESIS – Leibniz-Institute for the Social Sciences, Mannheim, Germany*. https://doi.org/10.15465/gesis-sg_036

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence, 4*(4), 14.

Lugtig, P., & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review, 34*(1), 78–94.

Lyberg, L. (2011). *The Paradata Concept in Survey Research*. https://csdiworkshop.org/wp-content/uploads/2020/03/Lybert2011CSDI.pdf. Presented at CSDI Workshop in London, UK, March 24, 2011. Pdf Accessed Jun 24, 2022.

Lynn, P. (2003). PEDAKSI: Methodology for collecting data about survey non-respondents. *Quality & Quantity, 37*(3), 239–261.

Maitland, A., & Presser, S. (2018). How do question evaluation methods compare in predicting problems observed in typical survey conditions? *Journal of Survey Statistics and Methodology, 6*(4), 465–490.

Matjašič, M., Vehovar, V., & Manfreda, K. L. (2018). Web survey paradata on response time outliers: A systematic literature review. *Advances in Methodology and Statistics (Metodološki zvezki), 15*(1), 23–41.

Matthijsse, S. M., De Leeuw, E. D., & Hox, J. J. (2015). Internet panels, professional respondents, and data quality. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 11*(3), 81–88.

Mayerl, J. (2013). Response latency measurement in surveys. Detecting strong attitudes and response effects. *Survey Methods: Insights From the Field, 27*, 1–26.

Mayerl, J., Sellke, P., & Urban, D. (2005). *Analyzing cognitive processes in CATI-Surveys with response latencies: An empirical evaluation of the consequences of using different baseline speed measures*. Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart -SISS- (Vol. 2/2005). Universität Stuttgart, Fak. 10 Wirtschafts- und Sozialwissenschaften, Institut für Sozialwissenschaften, Stuttgart, Germany. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-117346

McClain, C. A., Couper, M. P., Hupp, A. L., Keusch, F., Peterson, G., Piskorowski, A. D., & West, B. T. (2019). A typology of web survey paradata for assessing total survey error. *Social Science Computer Review, 37*(2), 196–213.

McLauchlan, C., & Schonlau, M. (2016). Are final comments in web survey panels associated with next-wave attrition? *Survey Research Methods, 10*(3), 211–224.

Meitinger, K., Ackermann-Piek, D., Blohm, M., Edwards, B., Gummer, T., & Silber, H. (2020). Special Issue: Fieldwork Monitoring Strategies for Interviewer-Administered Surveys. *Survey Methods: Insights from the Field*. https://core.ac.uk/download/pdf/343333745.pdf, https://surveyinsights.org/?p=13732

Minderop, I., & Weiß, B. (2023). Now, later, or never? Using response-time patterns to predict panel attrition. *International Journal of Social Research Methodology, 26*(6), 693–706. Published online first.

Mittereder, F. K. (2019). *Predicting and Preventing Breakoff in Web Surveys*. Dissertation, University of Michigan, Ann Arbor, MI. https://deepblue.lib.umich.edu/handle/2027.42/149963

Mittereder, F. K., & West, B. T. (2022). A Dynamic survival modeling approach to the prediction of web survey breakoff. *Journal of Survey Statistics and Methodology, 10*, 979–1004.

Mohadjer, L., & Edwards, B. (2018). Paradata and dashboards in PIAAC. *Quality Assurance in Education, 26*(2), 263–277.

Mohler, P. P., Pennell, B.-E., & Hubbard, F. (2012). Survey documentation: Toward professional knowledge management in sample surveys. In E. D. De Leeuw, J. Hox, & D. Dillman (Eds.), *International handbook of survey methodology* (pp. 403–420). Routledge.

National Academies of Sciences, Engineering, and Medicine (NAS) (2022). *Transparency in statistical information for the national center for science and engineering statistics and all federal statistical agencies*. The National Academies Press. https://doi.org/10.17226/26360

Nicolaas, G. (2011). Survey paradata: A review. *Discussion Paper NCRM/017, ESRC National Centre for Research Methods Review paper*. https://eprints.ncrm.ac.uk/id/eprint/1719

Olbrich, L., Beste, J., Sakshaug, J. W., & Schwanhäuser, S. (2022). *The Influence of Audio Recordings on Interviewer Behavior*. Poster Presented at LMU Munich Department of Statistics Summer Retreat, 2022/07/08-09.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly, 70*(5), 737–758.

Olson, K. (2013). Paradata for nonresponse adjustment. *The Annals of the American Academy of Political and Social Science, 645*(1), 142–170.

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 43–72). Wiley.

Peytchev, A. (2012). Multiple imputation for unit nonresponse and measurement error. *Public Opinion Quarterly, 76*(2), 214–237.

Plewis, I., Calderwood, L., & Mostafa, T. (2017). Can interviewer observations of the interview predict future response? *Methods, Data, Analyses, 11*(1), 1–16.

Purdam, K., Sakshaug, J. W., Bourne, M., & Bayliss, D. (2020). Understanding ‘Don’t Know’ answers to survey questions – An international comparative analysis using interview paradata. *Innovation: The European Journal of Social Science Research*, 1–23. https://www.tandfonline.com/doi/abs/10.1080/13511610.2020.1752631

Revilla, M., & Ochoa, C. (2015). What are the links in a web survey among response time, quality, and auto-evaluation of the efforts done? *Social Science Computer Review, 33*(1), 97–114.

Roßmann, J., & Gummer, T. (2016). Using paradata to predict and correct for panel attrition. *Social Science Computer Review, 34*(3), 312–332.

Sakshaug, J. W. (2013). Using paradata to study response to within-survey requests. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 171–190). Wiley.

Sakshaug, J. W., & Kreuter, F. (2011). Using paradata and other auxiliary data to examine mode switch nonresponse in a "Recruit-and-Switch" telephone survey. *Journal of Official Statistics, 27*(2), 339–357.

Sakshaug, J. W., & Struminskaya, B. (2023). *Call for Papers: Augmenting Surveys with Paradata, Administrative Data, and Contextual Data. A Special Issue of Public Opinion Quarterly*. https://academic.oup.com/poq/pages/call-for-papers-augmenting-surveys

Sana, M., & Weinreb, A. A. (2008). Insiders, outsiders, and the editing of inconsistent survey data. *Sociological Methods & Research, 36*(4), 515–541.

Scheuren, F. (2001). Macro and micro paradata for survey assessment. In T. Black, K. Finegold, A. B. Garrett, A. Safir, F. Scheuren, K. Wang, & D. Wissoker (Eds.), *1999 NSAF Collection of Papers*, pages 2C–1–2C–15. Urban Institute. https://www.urban.org/sites/default/files/publication/61596/410138---NSAF-Collection-of-Papers.PDF

Schlosser, S., & Höhne, J. K. (2020). *ECSP – Embedded Client Side Paradata*. Note: the 2020 version is an expansion of the 2016 and 2018 versions. https://doi.org/10.5281/zenodo.3782592

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., & Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through r-indicators and partial R-indicators. *International Statistical Review, 80*(3), 382–399.

Schouten, B., Peytchev, A., & Wagner, J. R. (2017). *Adaptive survey design*. CRC Press.

Schwanhäuser, S., Sakshaug, J. W., & Kosyakova, Y. (2022). How to catch a falsifier: Comparison of statistical detection methods for interviewer falsification. *Public Opinion Quarterly, 86*(1), 51–81.

Schwarz, H., Revilla, M., & Struminskaya, B. (2022). Do previous survey experience and participating due to an incentive affect response quality? Evidence from the CRONOS panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 185*, 1–23.

Sendelbah, A., Vehovar, V., Slavec, A., & Petrovčič, A. (2016). Investigating respondent multitasking in web surveys using paradata. *Computers in Human Behavior, 55*, 777–787.

Shlomo, N. (2018). Statistical disclosure limitation: New directions and challenges. *Journal of Privacy and Confidentiality, 8*(1). https://journalprivacyconfidentiality.org/index.php/jpc/article/view/684

Sinibaldi, J., Trappmann, M., & Kreuter, F. (2014). Which is the better investment for nonresponse adjustment: Purchasing commercial auxiliary data or collecting interviewer observations? *Public Opinion Quarterly, 78*(2), 440–473.

Smith, T. W. (2011). The report of the international workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. *International Journal of Public Opinion Research, 23*(3), 389–402.

Stern, M. J. (2008). The use of client-side paradata in analyzing the effects of visual layout on changing responses in web surveys. *Field Methods, 20*(4), 377–398.

Sturgis, P., Maslovskaya, O., Durrant, G., & Brunton-Smith, I. (2021). The interviewer contribution to variability in response times in face-to-face interview surveys. *Journal of Survey Statistics and Methodology, 9*(4), 701–721.

Sun, H., Conrad, F. G., & Kreuter, F. (2021). The relationship between interviewer-respondent rapport and data quality. *Journal of Survey Statistics and Methodology, 9*(3), 429–448.

Tourangeau, R. (2021). Science and survey management. *Survey Methodology, 47*(1), 3–29.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883.

Vardigan, M., Granda, P. A., & Hoelter, L. F. (2016). Documenting survey data across the life cycle. In C. Wolf, D. Joye, T. W. Smith, & Y.-c. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 443–459). SAGE.

Wagner, J. R. (2008). *Adaptive Survey Design to Reduce Nonresponse Bias*. Dissertation, University of Michigan, Ann Arbor, MI, 2008. https://deepblue.lib.umich.edu/handle/2027.42/60831

Wagner, J. R. (2013a). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods, 7*(1), 45–55.

Wagner, J. R. (2013b). Using paradata-driven models to improve contact rates in telephone and face-to-face surveys. In F. Kreuter (Ed.), *Improving surveys with paradata: analytic uses of process information* (pp. 145–170). Wiley.

Wagner, J. R. (2019). Estimation of survey cost parameters using paradata. *Survey Practice, 12*(1).

Wagner, J. R., Olson, K., & Edgar, M. (2017). The utility of GPS data in assessing interviewer travel behavior and errors in level-of-effort paradata. *Survey Research Methods, 11*(3), 218–233.

Wagner, J. R., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G., & Ndiaye, S. K. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics, 28*(4), 477–499.

West, B. T. (2011). Paradata in survey research. *Survey Practice, 4*(4), 1–8.

West, B. T. (2013a). The effects of error in paradata on weighting class adjustments: A simulation study. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 361–388). Wiley.

West, B. T. (2013b). An examination of the quality and utility of interviewer observations in the national survey of family growth. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 176*(1), 211–225.

West, B. T. (2018a). Collecting interviewer observations to augment survey data. In D. L. Vannette & J. A. Krosnick (Eds.), *The palgrave handbook of survey research* (pp. 211–215). Palgrave Macmillan/Springer.

West, B. T. (2018b). Interviewer observations. In D. L. Vannette & J. A. Krosnick (Eds.), *The palgrave handbook of survey research* (pp. 537–548). Palgrave Macmillan/Springer.

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology, 5*(2), 175–211.

West, B. T., & Groves, R. M. (2013). A propensity-adjusted interviewer performance indicator. *Public Opinion Quarterly, 77*(1), 352–374.

West, B. T., & Li, D. (2019). Sources of variance in the accuracy of interviewer observations. *Sociological Methods & Research, 48*(3), 485–533.

West, B. T., & Sinibaldi, J. (2013). The quality of paradata: A literature review. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 339–359). Wiley.

West, B. T., & Trappmann, M. (2019). Effective strategies for recording interviewer observations: Evidence from the PASS study in Germany. *Survey Methods: Insights from the Field*.

West, B. T., Wagner, J. R., Coffey, S., & Elliott, M. R. (2023). Deriving priors for Bayesian prediction of daily response propensity in responsive survey design: Historical data analysis versus literature review. *Journal of Survey Statistics and Methodology, 11*(2), 367–392.

Wilkinson, L. R., Ferraro, K. F., & Kemp, B. R. (2017). Contextualization of survey data: What do we gain and does it matter? *Research in Human Development, 14*(3), 234–252.

Wulff, D. U., Kieslich, P. J., Henninger, F., Haslbeck, J., & Schulte-Mecklenbeck, M. (2021). Movement tracking of cognitive processes: A tutorial using mousetrap. Preprint. *PsyArxiv*. https://doi.org/10.31234/osf.io/v685r

Yan, T. (2021). Consequences of asking sensitive questions in surveys. *Annual Review of Statistics and Its Application, 8*, 109–127.

Yan, T., & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 73–96). Wiley.

Patrick Oliver Schenk is a statistician in the working group of Frauke Kreuter at the Department of Statistics at the Ludwig-Maximilians-University, Munich, Germany. His research in survey methodology is focused on paradata, particularly their potential to predict and to improve data quality. Simone Reuß and Patrick Oliver Schenk have been collaborating on bringing together statistics and archeology in research and teaching, co-winning their university's teaching innovation prize.

Simone Reuß is a Prehistoric archaeologist at the Ludwig-Maximilians-University, Munich, Germany, focusing on burial rites of the Urnfield Period in Central Europe and mobility of people, objects, and ideas. Another interest is the answering of archeological questions with the help of statistical methods. Simone Reuß and Patrick Oliver Schenk have been collaborating on bringing together statistics and archeology in research and teaching, co-winning their university's teaching innovation prize.

Making Research Code Useful Paradata

Richèl J. C. Bilderbeek

Abstract

Paradata is data about the data collection process that allows use and reuse of data. Within the context of computational research, computer code is the paradata of an experiment, allowing the study to be reproduced. A recent study recommended how to make paradata (more) useful, for paradata in general. This study applies those recommendations to computer code, using the field of genetic epidemiology as an example. The chapter concludes by some rules how to better code to serve as paradata, and hence allowing computational research to be more reproducible.

1 Introduction

Talk is cheap. Show me the code.

Linus Torvalds, 2000-08-25

Two different researchers in genetic epidemiology (more on that field later) write two equally good manuscripts that describe an experiment with computational steps. Both manuscripts are accepted by an equally prestigious journal after peer review. One researcher, however, does not supply the computer code (from now on: ‘code’) that was used to generate the results, where the other does. Are the conclusion of these papers to be trusted equally? Is this difference relevant and worth the effort? How common is it to share code, and, if shared, how can it be preserved? This

R. J. C. Bilderbeek (✉)
National Bioinformatics Infrastructure Sweden (NBIS), Uppsala, Sweden

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_3

chapter discusses why code, a type of paradata, should be supplied and what features it needs to have for it to be useful.

The concept of paradata (although not named as such yet) was introduced by Couper and colleagues, who developed a computer-assisted interview program that, among others, records all key strokes, measures the time used to answer each question, and even the time the monitor is turned off (Couper, 1998). The goal in that context was to assess and improve survey quality.

There exists no standard definition of paradata (Nicolaas, 2011; Sköld et al., 2022; Huvila, 2022). Without declaring it a formal definition, however, paradata can be understood in general sense as 'data about processes' (with e.g. survey paradata termed alternatively as records and audit trails) (Nicolaas, 2011), or in more specific sense in relation to data collection, as 'data about the data collection process' (Choumert-Nkolo et al., 2019). This chapter uses the latter as a working definition.

The code used in computational experiments is just that 'data about the data collection process' as it commonly downloads data, selects relevant subsets in those data, performs statistical tests, and generates figures. In each case, code answers the question: 'Where is it (i.e. the result) coming from?'. Code, hence, is paradata, and this chapter explores and illustrates the consequence of that premise.

A recent paper explores the use of paradata to increase the impact of data, stating that lack of paradata can be seen as 'a drastic constraint' in the use of data and offer some suggestions to make paradata useful for data re(use): Paradata should be comprehensive, documented in a useful way, the documentation and data should have co-evolved, and the paradata should be computer-friendly (Huvila, 2022). At first glance, these suggestions appear to work well for code and will be discussed in detail below.

There are multiple reasons why useful paradata matters. Most obvious is that having useful paradata gives an understanding of how data is produced. This knowledge helps researchers from different fields to understand each other and collaborate. Additionally, useful open data is needed for Open Science to convincingly show its benefits. Finally, in computational fields, it can help understand how scholarly knowledge is produced (Huvila, 2022).

This chapter discusses these general reasons applied to code and its implications for knowledge management, including recommendations on how to make code useful paradata in Sect. 6 (Fig. 1).

2 Code Availability

To determine how useful code is, it needs to be available. However, it is not common to publish the code of an experiment or analysis (Stodden, 2011; Read et al., 2015) (with a pleasant exception being Conesa & Beck, 2019). For example, in computer graphics, a field intimately familiar with computer code, 42% of 454 SIGGRAPH papers supply computer code (Bonneel et al., 2020). Another study analysed the reproducibility of registered reports in the field of psychology, where 60% of 62

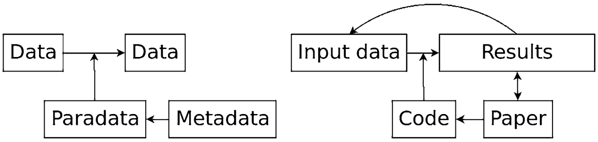


Fig. 1 **Left**: general relation between data, paradata, and metadata. **Right**: the same relations specified for genetic epidemiology. Input data: genetic data, phenotypic data, and associations found in earlier studies. Results: associations between genetic data and phenotypic data. Code: the computer code used in a computational experiment. Paper: the scholarly paper describing the (code of the) experiment

studies supplied the code to redo the analysis (Obels et al., 2020). For articles in Science magazine, 12% of 204 studies published the code (Stodden et al., 2018) (note that these were in years 2011–2012). Unpublished code has been the cause of some saddening examples, such as an algorithm that detects breast cancer from images better than a human expert, yet failed to ever be reproduced (Haibe-Kains et al., 2020).

When code of an academic paper is not published, one could contact the corresponding author and request it. However, the response rate of corresponding authors with a request for data in computational fields is around 50% (Manca et al., 2018; Stodden et al., 2018; Teunis et al., 2015) (the field of emergency medicine seems to be a pleasant outlier, with 73% of 118 emails being replied O'Leary, 2003). When getting an answer of the corresponding author, 48% out of 134 replied emails will actually result in a sharing of the code (Stodden et al., 2018). The responses of unwilling authors (see Stodden et al., 2018 for some real examples) can come across as so caustic that one may be excused from not contacting a corresponding author.

3 Genetic Epidemiology

This chapter uses genetic epidemiology as a specific example, to illustrate in which way code is paradata, and why this type of paradata is relevant. However, any field that uses computation and sensitive data in its experiments could be used as an example.

Genetic epidemiology is a field within biology that tries to determine the spread of heritable traits and their underlying biological mechanism. For example, we know that lactose intolerance in adults is caused by a decline in the production of lactose-degrading enzymes and is most commonly found in south-east Asia and south Africa (Storhaug et al., 2017). The trait is caused by the genetic make-up, or 'genotype' , of a person. The trait, also called 'phenotype', in this example is lactose intolerance at adult age, yet any human property, such as weight or height, can be studied. When an association between genotype and phenotype is found, these associations, when relevant enough, are used to create the so-called gene panels, where the location of the gene causing an association is measured specifically, to detect people at risk for the associated phenotype. The rest of this section describes a genetic epidemiology study in more detail, with special focus on the computational experiment.

The example study followed is a pseudorandomly selected paper from Ahsan et al. (2017). The primary data used by that paper is from a population study called the Northern Swedish Population Health Study (NSPHS) that started in 2010 (Igl et al., 2010). The approximately 1000 participants were initially mostly surveyed about lifestyle (Igl et al., 2010), and follow-up studies provided the type of data relevant for this paper, which are (1) the genotypes (Johansson et al., 2013), (2) the phenotypes, in this case, concentrations of certain proteins in the blood (Enroth et al., 2014, 2015).

The first type of primary data, the genotypes, consists of single nucleotide polymorphisms (SNPs, pronounced 'snips'). An SNP has a name and a location within the genome. At the SNP's location in the DNA, there will be two nucleotides. One of these nucleotides is inherited from the mother, the other from the father. The DNA consists of billions of nucleotides. There are four types of nucleotides, called adenosine, cytosine, guanine, and thyrosine, commonly abbreviated as A, C, G, and T, respectively.

One SNP example is `rs12133641`, which is an SNP located at position 154,428,283 (that is at the 154 millionth nucleotide), where 67% of the people within this study have an A, and 33% have a G (also from Ahsan et al., 2017, Table S3). From this it follows (assuming the nucleotides are inherited independently) that 45% of subjects have the genotype AA, 44% have AG, and 11% have GG.

The second type of data, the phenotypes, are concentrations of proteins in the blood. The nucleotides of the DNA contain the code for building proteins , as well as the rate at which a protein is created (for sake of simplicity, it is assumed that such a rate is constant, yet, in practice, there are complex regulation mechanisms). Some proteins end up in the blood, and their presence can be used to assess the health of an individual. IL6RA is one such protein, and its concentration may (and will, see below) be associated with the SNP mentioned earlier.

The field of genetic epidemiology looks—among others—for correlations between genetic data and biological traits. For example, Ahsan and colleagues show that SNP `rs12133641` is highly correlated (p-value is 3.0^{-73} , for n = 961 individuals) with protein IL6RA (Ahsan et al., 2017). The direction of the association is also concluded: The more guanines are present at that SNPs location, the higher concentration of IL6RA can be found in a human's blood . The amount of variance that can be explained by an association (i.e. the R^2) is rarely 100%, which means that a trait (in this case, the concentration of IL6RA) cannot be perfectly explained from the genotype (in this case, SNP `rs12133641`) alone. In this example, as much as 43% of the variance can be attributed to an individuals' genotype . Additional factors, such as the effect of the environment (e.g. geographic location, time of day the measurement was done), lifestyle (e.g. smoking yes/no), or having a disease (e.g. diabetes), are needed to explain the additional variation.

The conclusions drawn from this chapter may end up in the clinic. For the sake of having a clear (yet fictitious) example, let us assume that a high level of IL6RA is associated with a disease that develops later in life, yet is preventable by lifestyle changes (see Pope et al., 2003 for an example in Alzheimer's disease). Would this be the case, we can create a tailored experiment, called a gene panel, that specifically measures SNP `rs12133641`. If the gene panel shows an individual has two guanines, we know that this person is likelier to develop higher levels of IL6RA and is likelier to benefit from the lifestyle changes.

From this simple example, it will be easier to measure the level of IL6RA in the blood than using a gene panel, as blood tests are easier and cheaper. However, there are associations published for many diseases, in which one SNP (e.g. phenylketonuria) or many SNPs (e.g. Bruce & Byrne, 2009) contribute to being more likely to develop a disease in the future. Here, the phenotype (having a disease in the future) is impossible to detect at the present, and associations found in earlier studies are used to create a gene panel. As creating a gene panel is costly, those associations better be correct.

Additionally, there is an interdependency of scholarly findings here: The SNP has received its name based on a computational experiment. This earlier experiment that concluded the usefulness of that SNP is based on some DNA sequences. This experiment is based on assumed DNA sequences. DNA sequences, however, are (nowadays) not simply read, yet are the result of a complex computational analysis instead, with its own dedicated field of research. Both studies assumed a correctly calculated DNA sequence. This means that if the DNA sequence analysis contained a software bug, this study may be invalidated. Additionally, the result of this study may be used in follow-up studies that assume the result to be correctly calculated: One paper's conclusion is the next paper's assumption.

4 Why Code Is Useful Paradata

The experiment described above is run by code. It was code that detected the relationship between the genotype (in this case, SNP `rs12133641`) and the phenotype (in this case, the concentration of IL6RA). To be more precise, it was code that read the data, subsetted the data, removed outliers, performed the statistics, and generated the plots. For the rest of the discussion, we assume that the code is available to us (if not, see Sect. 2 for a glum estimate of the chance of obtaining the code).

There are multiple reasons why (useful) code matters, and these are the same reasons as why useful paradata matters: Code gives an understanding of how the raw results and the subsequent scholarly knowledge are obtained from an experiment. Additionally, code helps researchers from different fields to understand each other and collaborate. Additionally, code helps Open Science reach its goals of openness and transparency. The core of these reasons is to achieve reproducible science: That any person in any field can redo a computational experiment and see exactly what happened.

For computational science, it may appear to be relatively easy to reproduce an experiment, as all it takes is a computer, electricity, an optional Internet connection, the code, and the data. In practice, however, only 18% of 180 computational studies are easily reproducible (Stodden et al., 2018). To some, it appears that the academic culture to reproduce results has been lost over time (Peng, 2011), with labs that embrace reproducibility (for example Barba, 2016) being the exception. One suggested way forward is to make the reproduction of research a minimal requirement for publication (Peng, 2011).

A genetic epidemiologist works with sensitive data as well: The genetic sequences of participants are private Clayton et al., 2019. For research to be reproducible, one needs both the code and the data to reproduce the (hopefully) same results. This problem is discussed in Sect. 7.

Code holds the ground truth of an experiment; it does the actual work. The more complex the computation pipeline is, the easier it is to have a mismatch between the article (that describes what the code does) and the code (that actually does the work). The moment these two disagree, it is the code that is true.

5 Preserving Code

Code is rarely preserved (Barnes, 2010). This section discusses the preservation of code for a short, medium, and long term.

5.1 Code Hosting

To preserve code for a short term, a code hosting website is a good first step. A code hosting website is a website where its users can create dedicated pages (called 'repositories') for a project, upload code, and interact with that code. There are multiple code hosting websites, with GitHub being the most popular one (Cosentino et al., 2017). The use of code hosting websites has increased strongly (Russell et al., 2018), accommodates collaboration (Perez-Riverol et al., 2016), and improves transparency (Gorgolewski & Poldrack, 2016), due to its inherent computer-friendliness. See Cosentino et al. (2017) for an extensive overview of research conducted on GitHub.

Hosted code commonly keeps a history of file changes. This means that when a change is made to the code, a new version is created. In the case that the change was harmful, one can go back to an earlier version and continue again from there. The version control system keeps track of who-did-what transparently. It is a general recommendation to put version control on all human-produced data (Wilson et al., 2014), as well as openly working on the code from the start (Jiménez et al., 2017). Half of the published code has such a version control system (Stodden et al., 2018).

The degradation of software is a known feature for nearly four decades. This is called 'bit rot' by Steele Jr. et al. (1983), or 'software collapse' by Hinsen (2019), in which software fails due to dependencies on other software. Using a code hosting website, which only passively stores code, ignores this problem.

5.2 Continuous Integration

To preserve code for a longer timespan, it needs to be embraced that software degrades (Beck, 2000). Continuous integration ('CI') allows one to verify if code still works and, if not, to be notified.

Some code hosts allow a user to trigger specialized code upon uploading a change, called a CI script. Such a CI script typically builds and tests the code. This practice is known to significantly increase the number of bugs exposed (Vasilescu et al., 2015) and increase the speed at which new features are added (Vasilescu et al., 2015). CI can be scheduled to run on a regular basis and notify the user directly when the code has broken down.

5.3 Containerization

To preserve code for the longest time, both code and its dependencies can be put into a so-called container. The most reproducible way of submitting the code of an experiment is to put all code with all the (software) dependencies in a file that acts as a virtual computational environment, called a 'virtual container' (from now on: 'container'). Such a container is close to the golden standard of reproducible research as suggested by Peng (2011) .

6 Making Code Useful Paradata

Useful paradata, in general, is (1) comprehensive, (2) documented appropriately, (3) documented in co-evolution with the data, and (4) friendly to computers (Huvila, 2022). In this section, these ideal properties of paradata are applied to code. However, code has multiple uses, as code can be used to (1) reproduce, (2) replicate, or (3) extend a computational experiment (Benureau & Rougier, 2018). Depending on the intended use of the code, there are different requirements for code being useful paradata.

6.1 Code Must Be Usefully Documented

For code to be ideal paradata, it must be usefully documented (Huvila, 2022). For purposes of reproducing code, it should at least be documented how to run the code and what it ought to do. Although this may be obvious, only 57% out of 56 science papers with obtainable code (in total there were 180 papers) do so (Stodden et al., 2018).

When it can be found out how an experiment is run, it is possible (even ideal!) that no code is read at all. Within that context, it could be argued that no (further) documentation is needed. However, all code in general should be documented 'adequately' (Peng et al., 2006), ideally writing code in such a way that it becomes self-explanatory (Wilson et al., 2014) and for the remaining code to document the reasons behind it, its design, and its purpose (Wilson et al., 2014).

For purposes of extending a study and its code, documentation becomes even more important, as the code will be read and modified. The extent of investing time in documenting code is recommended to be proportional to the intended reuse (Pianosi et al., 2020), and there exists a clear relationship between the reuse of code and its documentation effort (Cosentino et al., 2017; Hata et al., 2015).

6.2 Code and Documentation Must Align

For code to be ideal paradata, its creation and documentation need to align, *not least because they shape each other* (Huvila, 2022). This emphasized part of the quote resonates strongly with the idea of literate programming (Knuth, 1984), in which documentation and code are developed hand-in-hand. Literate programming is the practice of writing code and documentation in the same file. Contemporary examples of this idea are, among others, vignettes (Wickham, 2015) for the R programming language (R Core Team, 2021) and Jupyter notebooks (Wang et al., 2020) for the Julia (Bezanson et al., 2017), Python (Van Rossum & Drake Jr., 1995), and R (R Core Team, 2021) programming languages. rOpenSci, a community that, among others, reviews programming code (Ram, 2013; Ram et al., 2018), is one example of where extensive documentation is mandatory and all code must have examples (that are actually run) as part of the documentation (rOpenSci et al., 2021). In general, when developing software, it is recommended to to write documentation while writing software, as well as to include many examples (Lee, 2018), as this leads to both better code and documentation (Reenskaug & Skaar, 1989).

6.3 Code Must Be Extensive

For code to be ideal paradata, it must be extensive. As code has many properties, there are many recommendations on this aspect.

Code should be distributed in standard ways (Peng et al., 2006), as is done by using a code hosting website (see Sect. 5.1). Additionally, code must be more extensive when it is (intended to be) used on different data, as then 'code must act as a teacher for future developers' (Sadowski et al., 2018). Error handling is one of the mechanisms to do so. In genetic epidemiology, it is common to have incomplete or missing data, so analyses should take this into account with clear error messages.

Coding errors are extremely common (Baggerly & Coombes, 2009; Vable et al., 2021) and contribute to the reproducibility crisis in science (Vable et al., 2021). Testing, in general, is an important mechanism to ensure the correctness of code. One clear example is Rahman and Farhana (2020), showing bugs in scientific software on the COVID-19 pandemic.

Testing is so important that it is at the heart of a software development methodology called ‘Test-Driven Development’ (‘TDD’), in which tests are written before the ‘real’ code. TDD improves code quality (Alkaoud & Walcott, 2018; Janzen & Saiedian, 2006), and it is easy to integrate the writing of documentation as part of the TDD cycle (Shmerlin et al., 2015).

The percentage of (lines of) code tested is called the code coverage. Code coverage correlates with code quality (Horgan et al., 1994; Del Frate et al., 1995), and, due to this, having a code coverage of (around) 100% is mandatory to pass a code peer review by rOpenSci (Ram et al., 2018). When CI is activated, the code coverage of a project can be shown on the repository’s website .

It is considered good practice to add a software license (Jiménez et al., 2017), so that it is clear that the software can be reused. Although this may seem trivial, only two-thirds of 56 computational experiments supply a software license (Stodden et al., 2018).

Code reviews are recommended by software development best practices (Wilson et al., 2014). However, more than half of 315 scientists have their code rarely or never reviewed (Vable et al., 2021), although code reviews are known to accelerate learning of the developers, improve the quality of the code, and resulting to an experiment that is likelier to be reproducible (Vable et al., 2021).

6.4 Code Must Be Computer-Friendly

The most reproducible way of submitting the code of an experiment is by providing the code with all its (software) dependencies in a container. Containers allow a computation experiment to be highly reproducible: Given the same data, an experiment put into a container will give the same results on different platforms, at least in theory. In practice, differences may be observed when peripheral factors are different, such as the random numbers as generated by an operating system, or data that are downloaded from online (and hence, probably changing) sources.

For paradata to be useful, it has to be computer-friendly, yet “the best paradata does not necessarily look like ‘data’ at all for its human users” (Huvila, 2022). There are features of code that humans find useful, without directly being able to measure these. In the end, code is just ‘another kind of data’ and should be designed as such, for example by using tools to work on it (Wilson, 2022).

A first example is to use a tool to enforce a coding style [e.g. the Tidyverse style guide (Wickham, 2019) for R, or PEP 8 (Van Rossum et al., 2001) for Python], as following a consistent coding style improves software quality (Fang, 2001). A second example is to use a tool to enforce a low cyclomatic complexity. The cyclomatic complexity is approximately defined as the number of independent paths that the code can be executed. The cyclomatic complexity correlates with code complexity, where more complex code is likelier to contain or give rise to bugs (Abd Jader & Mahmood, 2018; Chen, 2019; Zimmermann et al., 2008).

7 Sensitive Data

Next to the code, it is the data used in an experiment that must be made available for an experiment to be called ‘reproducible’ (Peng et al., 2006). In some fields, such as genetic epidemiology, the data is sensitive, hence cannot be released, and thus one cannot reproduce an experiment. To solve this problem in the future, there are some interesting methods being developed to run code on sensitive data with assured privacy (Zhang et al., 2016; Azencott, 2018).

To alleviate the problem today, a developer should supply a simulated [also called ‘analytical’ (Peng et al., 2006)] dataset together with the code. This simulated data is needed to run tests, as is part of the TDD methodology. In the case of genetic epidemiology, this would mean simulated genotypes and associated phenotypes, as can be done with the `plinkr` R package (Bilderbeek, 2022). One extra benefit of simulated data is that these can be used as a benchmark, as slightly different analyses should give similar conclusions.

8 Discussion

In a perfect world, all code has the characteristics of ideal paradata and is written from software development best practices. This section discusses the problems that arise by doing so.

To know these best practices, one needs to be trained. Articles that suggest these best practices (such as this one) claim that this initial investment pays off. Code reviews are a good way to accelerate the learning of team members (Vable et al., 2021).

Code needs maintenance, as code that will stand the test of time perfectly is deemed ‘impossible’ (Benureau & Rougier, 2018). CI can help a maintainer to be notified when the code breaks, where the use of containers may slow down time, as an entire computational environment is preserved.

Uploading code, preferably to a code hosting website, may feel like a risk, as all code can be seen and scrutinized. However, not publishing code may put oneself in the focus of attention and—after much effect by others reproducing an incorrect result— at the cost of a scientific career (Baggerly & Coombes, 2009).

When the author of code can be contacted, there will be users asking for technical support. One solution for the author is to ignore such emails, as is done in a third of 357 cases (Teunis et al., 2015): It can be argued that no energy should be wasted on published code and work on something new instead. However, see Barnes (2010) for a better way to deal with this problem.

When the author of code can be contacted, users will send in bug reports. If the bug is severe enough, the question arises if all the research that use that code still results in the same conclusions. One such bug is described in Eklund et al. (2016), with 40,000 studies using that incorrect code. These reports could be ignored to work on something new.

Containers do have problems. First, they themselves require software to run, with the same software decay being possible. Second, one needs to install that software and have the computer access rights (i.e. admin rights under Windows, or root rights under Linux) to do so. Third, one needs to learn how to build and use containers. Lastly, containers can be several gigabytes big files, which makes their distribution harder. Ideally, containers are stored online and distributed in standardized ways. Although progress is being made, there is no way to do so for all container types. Additionally, probably due to their novelty, container hosting sites lack metadata.

9 Conclusions

This chapter started with some suggestions to make paradata useful for data re(use): Paradata should be extensive, comprehensively documented, with the creation of documentation and code going hand-in-hand, as well as friendly to computers (Huvila, 2022).

Before applying these features, the first step is to publish the code. When applying these general recommendations to code, this list can be phrased more precisely:

1. Code should be comprehensive in supplying automatically generated metadata (such as commit history and code coverage).
2. The documentation should be as extensive as recommended by the software development literature.
3. The documentation should have co-evolved with the code following the best practices in literate programming.
4. Code should be made machine-readable by, at least, being uploaded to a code hosting website. Ideally, the code is checked by CI and is put in a container.

For the preservation of code, these recommendations are made:

1. Uploading code to a code hosting website is better than not publishing code at all.
2. Adding CI to code allows one to detect the day when that code does not run anymore.
3. Putting the code in a container is the best way to preserve code.

When research truly needs to be reproducible, putting the code of an experiment into a container is today’s best solution, as containers are the best solution to keep code running for the longest amount of time. Creating such a container, however, requires more skill that—as of today—is not rewarded, although an experiment put into a container can be considered the pinnacle of reproducible research.

The simplest and most impactful step to make code more useful paradata is, however, to publish it on a code hosting website along a publication. From then on, the next steps can be taken gradually as the skills of the author(s) progress. To quote Barnes, 2010: Publish your code, and it is good enough.

The world of science would be a more open, humble, trustworthy, truthful and helpful would the code that accompanies a scientific paper be given the attention it deserves treated like a first class citizen. Doing so, however, is yet to be rewarded, and still both of the two scientists at the start of this chapter can provide a good rationale for their behaviour. This will change when reward incentives are put into place that reward making paradata useful. For code specifically, in any computational field, the rewards are even higher, as reproducibility should again be a cornerstone in science.

10 Data Accessibility

This chapter and its metadata can be found at https://github.com/richelbilderbeek/chapter_paradata.

References

Abd Jader, M. N., & Mahmood, R. Z. (2018). Calculating McCabe's cyclomatic complexity metric and its effect on the quality aspects of software. *International Journal of Innovative Research and Creative Technology, 3*, 10–22.

Ahsan, M., Ek, W. E., Rask-Andersen, M., Karlsson, T., Allan Lind-Thomsen, Enroth, S., Gyllensten, U., & Johansson, Å. (2017). The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. *PLoS Genetics, 13*(9), e1007005.

Alkaoud, H., & Walcott, K. R. (2018). Quality metrics of test suites in test-driven designed applications. *International Journal of Software Engineering Applications (IJSEA), 9*, 1–16.

Azencott, C.-A. (2018). Machine learning and genomics: Precision medicine versus patient privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), 20170350.

Baggerly, K. A., & Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics, 3*, 1309–1334.

Barba, L. A. (2016). The hard road to reproducibility. *Science, 354*(6308), 142.

Barnes, N. (2010). Publish your computer code: It is good enough. *Nature, 467*(7317), 753.

Beck, K. (2000). *Extreme programming explained: Embrace change*. Addison-Wesley Professional.

Benureau, F. C. Y., & Rougier, N. P. (2018). Re-run, repeat, reproduce, reuse, replicate: Transforming code into scientific contributions. *Frontiers in Neuroinformatics, 11*, 69.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review, 59*(1), 65–98. https://doi.org/10.1137/141000671. https://epubs.siam.org/doi/10.1137/141000671

Bilderbeek, R. J. C. (2022). https://github.com/richelbilderbeek/plinkr. Accessed Augest 08, 2022.

Bonneel, N., Coeurjolly, D., Digne, J., & Mellado, N. (2020). Code replicability in computer graphics. *ACM Transactions on Graphics (TOG), 39*(4), 93–1.

Bruce, K. D., & Byrne, C. D. (2009). The metabolic syndrome: Common origins of a multifactorial disorder. *Postgraduate Medical Journal, 85*(1009), 614–621.

Chen, C. (2019). An empirical investigation of correlation between code complexity and bugs. *arXiv preprint arXiv:1912.01142*.

Choumert-Nkolo, J., Cust, H., & Taylor, C. (2019). Using paradata to collect better survey data: Evidence from a household survey in Tanzania. *Review of Development Economics, 23*(2), 598–618.

Clayton, E. W., Evans, B. J., Hazel, J. W., & Rothstein, M. A. (2019). The law of genetic privacy: Applications, implications, and limitations. *Journal of Law and the Biosciences, 6*(1), 1–36.

Conesa, A., & Beck, S. (2019). Making multi-omics data accessible to researchers. *Scientific Data, 6*(1), 1–4.

Cosentino, V., Izquierdo, J. L. C., & Cabot, J. (2017). A systematic mapping study of software development with GitHub. *IEEE Access, 5*, 7173–7192.

Couper, M. (1998). Measuring survey quality in a CASIC environment. In *Proceedings of the Survey Research Methods Section of the ASA at JSM1998* (pp. 41–49).

Del Frate, F., Garg, P., Mathur, A. P., & Pasquini, A. (1995). On the correlation between code coverage and software reliability. In *Proceedings., Sixth International Symposium on Software Reliability Engineering, 1995* (pp. 124–132). IEEE.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences, 113*(28), 7900–7905.

Enroth, S., Enroth, S. B., Johansson, Å., & Gyllensten, U. (2015). Effect of genetic and environmental factors on protein biomarkers for common non-communicable disease and use of personally normalized plasma protein profiles (PNPPP). *Biomarkers, 20*(6–7), 355–364.

Enroth, S., Johansson, Å., Enroth, S. B., & Gyllensten, U. (2014). Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nature Communications, 5*(1), 1–11.

Fang, X. (2001). Using a coding standard to improve program quality. In *Proceedings. Second Asia-Pacific Conference on Quality Software, 2001* (pp. 73–78). IEEE.

Gorgolewski, K. J., & Poldrack, R. (2016). A practical guide for improving transparency and reproducibility in neuroimaging research. *bioRxiv*, p. 039354.

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Board, M. S., Waldron, L., Wang, B., Mcintosh, C., Kundaje, A., Greene, C., et al. (2020). *The importance of transparency and reproducibility in artificial intelligence research.*

Hata, H., Todo, T., Onoue, S., & Matsumoto, K. (2015). Characteristics of sustainable OSS projects: A theoretical and empirical study. In *2015 IEEE/ACM 8th International Workshop on Cooperative and Human Aspects of Software Engineering* (pp. 15–21). IEEE.

Hinsen, K. (2019). Dealing with software collapse. *Computing in Science & Engineering, 21*(3), 104–108.

Horgan, J. R., London, S., & Lyu, M. R. (1994). Achieving software quality with testing coverage measures. *Computer, 27*(9), 60–69.

Huvila, I. (2022). Improving the usefulness of research data with better paradata. *Open Information Science, 6*(1), 28–48.

Igl, W., Johansson, Å., & Gyllensten, U. (2010). The Northern Swedish population health study (NSPHS)–a paradigmatic study in a rural population combining community health and basic research. *Rural and Remote Health, 10*(2), 198–215.

Janzen, D. S., & Saiedian, H. (2006). Test-driven learning: intrinsic integration of testing into the CS/SE curriculum. *Acm Sigcse Bulletin, 38*(1), 254–258.

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., Capella-Gutierrez, S., Hong, N. C., Cook, M., Corpas, M., et al. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research, 6*, ELIXIR-876.

Johansson, Å., Enroth, S., Palmblad, M., Deelder, A. M., Bergquist, J., & Gyllensten, U. (2013). Identification of genetic variants influencing the human plasma proteome. *Proceedings of the National Academy of Sciences, 110*(12), 4673–4678.

Knuth, D. E. (1984). Literate programming. *The Computer Journal, 27*(2), 97–111.

Lee, B. D. (2018). Ten simple rules for documenting scientific software. *PLOS Computational Biology, 14*(12), e1006561.

Manca, A., Cugusi, L., Dvir, Z., & Deriu, F. (2018). Non-corresponding authors in the era of meta-analyses. *Journal of Clinical Epidemiology, 98*, 159–161.

Nicolaas, G. (2011). *Survey paradata: a review*. National Centre for Research Methods.

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 229–237.

O'Leary, F. (2003). Is email a reliable means of contacting authors of previously published papers? A study of the emergency medicine journal for 2001. *Emergency Medicine Journal, 20*(4), 352–353.

Peng, R. D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226–1227.

Peng, R. D., Dominici, F., & Zeger, S. L. (2006). Reproducible epidemiologic research. *American Journal of Epidemiology, 163*(9), 783–789.

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F., Fufezan, C., Ternent, T., Eglen, S. J., Katz, D. S. S., et al. (2016) Ten simple rules for taking advantage of git and GitHub. *bioRxiv*, p. 048744.

Pianosi, F., Sarrazin, F., & Wagener, T. (2020). How successfully is open-source research software adopted? Results and implications of surveying the users of a sensitivity analysis toolbox. *Environmental Modelling & Software, 124*, 104579.

Pope, S. K., Shue, V. M., & Beck, C. (2003). Will a healthy lifestyle help prevent Alzheimer's disease? *Annual Review of Public Health, 24*(1), 111–132.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Rahman, A., & Farhana, E. (2020). An exploratory characterization of bugs in covid-19 software projects. *arXiv preprint arXiv:2006.00586*.

Ram, K. (2013). rOpenSci-open tools for open science. In *AGU Fall Meeting Abstracts* (Vol. 2013, pp. ED43E–04).

Ram, K., Boettiger, C., Chamberlain, S., Ross, N., Salmon, M., & Butland, S. (2018). A community of practice around peer review for long-term research software sustainability. *Computing in Science & Engineering, 21*(2), 59–65.

Read, K. B., Sheehan, J. R., Huerta, M. F., Knecht, L. S., Mork, J. G., Humphreys, B. L., & NIH Big Data Annotator Group (2015). Sizing the problem of improving discovery and access to NIH-funded data: a preliminary study. *PLoS One, 10*(7), e0132735.

Reenskaug, T., & Skaar, A. L. (1989). An environment for literate Smalltalk programming. In *Conference Proceedings on Object-Oriented Programming Systems, Languages and Applications* (pp. 337–345).

rOpenSci, Anderson, B., Chamberlain, S., DeCicco, L., Gustavsen, J., Krystalli, A., Lepore, M., Mullen, L., Ram, K., Ross, N., Salmon, M., Vidoni, M., Riederer, E., Sparks, A., & Hollister, J. (2021). *rOpenSci Packages: Development, Maintenance, and Peer Review*. https://doi.org/10.5281/zenodo.6619350

Russell, P. H., Johnson, R. L., Ananthan, S., Harnke, B., & Carlson, N. E. (2018). A large-scale analysis of bioinformatics code on GitHub. *PLoS One, 13*(10), e0205898.

Sadowski, C., Söderberg, E., Church, L., Sipko, M., & Bacchelli, A. (2018). Modern code review: a case study at Google. In *Proceedings of the 40th International Conference on Software Engineering: Software Engineering in Practice* (pp. 181–190).

Shmerlin, Y., Hadar, I., Kliger, D., & Makabee, H. (2015). To document or not to document? An exploratory study on developers' motivation to document code. In *Advanced Information Systems Engineering Workshops: CAiSE 2015 International Workshops, Stockholm, Sweden, June 8-9, 2015, Proceedings 27* (pp. 100–106). Springer.

Sköld, O., Börjesson, L., & Huvila, I. (2022). *Interrogating paradata*.

Steele Jr., G. L., Woods, D. R., Finkel, R. R., Stallman, R. M., & Goodfellow, G. S. (1983). *The hacker's dictionary: A guide to the world of computer wizards*. Harper & Row Publishers.

Stodden, V. C. (2011). *Trust your science? Open your data and code*.

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences, 115*(11), 2584–2589.

Storhaug, C. L., Fosse, S. K., & Fadnes, L. T. (2017). Country, regional, and global estimates for lactose malabsorption in adults: a systematic review and meta-analysis. *The Lancet Gastroenterology & Hepatology, 2*(10), 738–746.

Teunis, T., Nota, S. P. F. T., & Schwab, J. H. (2015). Do corresponding authors take responsibility for their work? A covert survey. *Clinical Orthopaedics and Related Research®, 473*, 729–735.

Vable, A. M., Diehl, S. F., & Glymour, M. M. (2021). Code review as a simple trick to enhance reproducibility, accelerate learning, and improve the quality of your team's research. *American Journal of Epidemiology, 190*(10), 2172–2177.

Van Rossum, G., & Drake Jr., F. L. (1995). *Python tutorial* (Vol. 620). Centrum voor Wiskunde en Informatica Amsterdam.

Van Rossum, G., Warsaw, B., & Coghlan, N. (2001). PEP 8–style guide for Python code. *Python. org, 1565*, 28.

Vasilescu, B., Yu, Y., Wang, H., Devanbu, P., & Filkov, V. (2015). Quality and productivity outcomes relating to continuous integration in GitHub. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering* (pp. 805–816). ACM.

Wang, J., Kuo, T.-Y., Li, L., & Zeller, A. (2020). Assessing and restoring reproducibility of Jupyter notebooks. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering* (pp. 138–149).

Wickham, H. (2015). *R packages: Organize, test, document, and share your code*. O'Reilly Media.

Wickham, H. (2019). *Advanced R*. CRC Press.

Wilson, G. (2022). Twelve quick tips for software design. *PLoS Computational Biology, 18*(2), e1009809.

Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., et al. (2014). Best practices for scientific computing. *PLoS Biology, 12*(1), e1001745.

Zhang, L., Zheng, Y., & Kantoa, R. (2016). A review of homomorphic encryption and its applications. In *Proceedings of the 9th EAI International Conference on Mobile Multimedia Communications* (pp. 97–106).

Zimmermann, T., Nagappan, N., & Zeller, A. (2008). Predicting bugs from history. In *Software evolution* (pp. 69–88). Springer.

Richèl J. C. Bilderbeek has worked in computational biology for more than a decade. Currently, he works for the National Bioinformatics Infrastructure Sweden (NBIS) in Uppsala, Sweden, as an application expert and trainer, with a focus on reproducible research. All his work can be found on GitHub at https://github.com/richelbilderbeek and his YouTube channel.

A Leap of Faith: Revisiting Paradata in 3D Scholarship

Costas Papadopoulos

Abstract

3D visualisation—be it computer graphic (re)construction or digitisation—has a long tradition in archaeology and cultural heritage; original research approaches, new methodologies, and theoretical frameworks have been developed; scholarly outputs in a range of forms have been published; teaching programmes have been designed; and an array of apparatuses, including organisations, consortia, projects, conferences, journals, and book series exclusively focusing on 3D, have been established. Despite all these, 3D scholarship is still faced with scepticism and hesitation, both due to constant changes in technology and the fragile ecosystem within which it is being developed, but also due to the technological authority, lack of standards, and its non-conventional nature that does not adhere to established academic norms. The development of charters and principles, such as the London and Seville Charters, which were developed to provide guidelines that tackle these issues, have been inconsistently addressed and rarely implemented. By looking back at the origins of paradata in heritage visualisation and the ways that three-dimensionality is perceived, captured, and interpreted in conventional archaeological practice, this chapter explores the variable and dialectic processes that take place at the trowel's edge and the often neglected perceptual, physiological, and technical factors that influence knowledge production in the process of 3D (re)construction. The chapter argues that 3D scholarship requires a leap of faith and a rethinking of the 'how, when, and why' of paradata on a par with our better understanding of the complexity of and recent changes in digital scholarship.

C. Papadopoulos (✉)
Department of Literature & Art, Faculty of Arts & Social Sciences, Maastricht University, Maastricht, The Netherlands
e-mail: k.papadopoulos@maastrichtuniversity.nl

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_4

1 Introduction

The concept of paradata has the potential to provide the basis for intellectual transparency (Sköld et al., 2022) by capturing information about the processes of understanding and interpretation leading to scholarly outputs. At the same time, it is a concept that has haunted (Dawson and Reilly, chapter 'Towards Embodied Paradata. A Diffractive Art/Archaeology Approach' in this volume) the field of 3D heritage visualisation since the inception of the London Charter (LC) in 2006. The LC, a set of principles to ensure the technical and intellectual transparency of 3D visualisation, came as a response to over a decade of intense debates about 3D scholarship in archaeology and cultural heritage. Advancements in computational hardware and rendering algorithms, and the ability to create photorealistic computer-generated images (CGI) that were indistinguishable from real life, made several scholars argue that 3D was becoming 'the downright misleading' (Miller & Richards, 1995) and 'a double-edged sword' (Eiteljorg, 2000). Therefore, solutions had to be found to safeguard scholars and their scholarship by means of ensuring that visualisations were not only providing engaging images for public consumption but could also serve and be accepted by a scholarly audience.

Another challenge that 3D scholarship had to face was the absence of standards and stable systems for documenting and presenting processes and outputs. Until recently, major changes in technologies and proprietary systems had as a result the loss of substantial amounts of 3D scholarship (Papadopoulos & Schreibman, 2019), at least in its original form. Considering these issues, the discussion about intellectual transparency and the documentation of processes and interpretations was a timely and worthwhile pursuit. The LC formalised this discussion under a series of principles that aimed at providing a foundation on which different communities of practice could develop a series of implementable rules. Ultimately, the LC aimed at legitimising a particular type of work which although had many practitioners, it did not have many researchers advocating for its value as an autonomous piece of scholarship.

The current hype for Virtual Reality and the Metaverse has brought—after a long hiatus—3D graphics to the fore. Also, a renewed interest has been evident in heritage research and dissemination with several projects problematising issues of sustainability, quality, intellectual transparency, and infrastructures.[1] In addition, there have been significant attempts to better integrate 3D into traditional scholarship; for example, journals such as Studies in Digital Heritage and *Digital Applications in Archaeology and Cultural Heritage* have been accepting and peer-reviewing 3D models as part of journal articles, while publishers, such as Stanford University Press and Michigan University Press, have been publishing digital

[1] See for example: Community Standards for 3D Data Preservation—Moore et al. (2022); PURE3D: An Infrastructure for the Presentation and Preservation of 3D Scholarship; Dynamic Collections—Callieri et al. (2020); Study on quality in 3D digitisation of tangible cultural heritage—European Commission (2022)

monographs that integrate 3D into their narratives (see, e.g., Opitz et al., 2018; Sullivan, 2020). Such developments have created a more open environment for 3D scholarship, which was not the case when the LC was conceived. The European Commission (2021) is in the process of developing through a series of dedicated Horizon Europe funding calls a Common European Data Space for cultural heritage which will also involve the 3D digitisation of all at risk and most visited cultural heritage. This, together with the increasing digitisation of heritage—mostly afforded by the changes in web technology and the democratisation of digitisation tools and methods—will inevitably create more challenges for 3D scholarship.

Is such mass digitisation a threat to the steps that have been already taken towards a more sustainable 3D scholarship or is it an opportunity to re-evaluate its role and meaning in our field? This may also provide us with a clean slate to critically discuss how should 3D models be presented on the Web; what kind of information should be captured about their making; and how should this be made available to the intended audience. None of these are insignificant undertakings and such decisions have an impact on understanding the outputs of this scholarship, especially when there is no access to the people, hardware, and software that developed and defined them. Even if such access was possible, the conscious and unconscious decisions that human operators made but also technological authority, i.e. decisions that software and hardware made for us, need to be considered. However, such decisions often remain blackboxed. Devising systems to capture as much information as possible in the hope that we can ultimately (re)construct every step of the process is not only impossible (Schreibman & Papadopoulos, 2019) but also futile. What purpose would such a complete documentation serve? Although transparency has essentially become the dominant narrative around 3D scholarship, one can only wonder if 3D scholars need to be more transparent than others. Are there fundamental and unique differences in their practice? Are processes and decisions objective and linear or are they defined by unknown variables? Ultimately, is the concept of paradata still relevant, and if so, how can we ensure that it aligns with current and future developments in (3D) digital scholarship?

Since 3D visualisations can be generated in different ways and by different means (Huvila, 2018), and each of those bears unique technological and epistemological challenges, this chapter will only problematise the modelling process, i.e. the development of a 3D model by using a computer graphics software. It will argue that before reaching the stage of formalising the documentation of processes and decisions, we ought to understand that 3D (re)construction is ‘unbearably complex’ (Huvila, 2013) since it involves variable and dialectic processes based on a series of often neglected perceptual, physiological, and technical factors. Since technological authority is only one of the facets that need to be examined, we cannot and should not treat 3D (re)constructions as merely technical or technological processes. The premise of this chapter is that in order to critically evaluate if paradata still have a role to play in such processes, if they need to be redefined, and/or if they ultimately become counterintuitive, elusive, or illusionary (Reilly et al., 2021), we should first problematise what comes before a 3D (re)construction, i.e. the complex processes of decision-making and knowledge production in archaeological practice.

To approach this question, the chapter is structured as follows. First, it provides an origin story of paradata in 3D heritage visualisation by exploring various charters, guidelines, and their implementations. Then, it turns into archaeological practice and examines the concept of three-dimensionality, particularly focusing on how it is perceived and documented. Lastly, it returns to the concept of paradata and discusses its relevance. The chapter concludes by arguing that paradata need to be reconsidered within an ecosystem that recognises and rewards non-typical scholarship. It proposes a 'leap of faith' to provide intellectual rigour and facilitate the transition to more contemporary conceptions of scholarship while escaping from any paradigms that are based on the notion of reproducibility.

2 Paradata: Tracing the Origins and Paving the Futures

Mick Couper coined the term paradata in 1988 to differentiate the data automatically generated by computer-assisted interviewing from those produced by humans in the survey process. As he outlines in a review article on the birth and development of the term (Couper, 2017), technological developments in the field, e.g. the growth of Web surveys and computer-based systems for survey research, led to the expansion of the concept to also include the description of contextual circumstances beyond the survey itself, e.g. observations, which may prove useful in the management, understanding, and evaluation of the collected data.

The term paradata follows a similar path in heritage visualisation. With the increasing computational capacity in the 2000s and the move from schematic, illustration-like 3D visualisations to photorealistic ones, researchers started questioning the validity and consequently the scholarly value of such representations, while exploring ways that the intellectual rigour that went into their creation, including accuracy, uncertainty, and interpretation, can be communicated. As a result, two separate solution-oriented directions were developed: a computational and a conceptual one. The two approaches did not develop in parallel; computational approaches were developed well before the term paradata was used in 3D heritage visualisation, while the conceptual one started developing with the symposium on *Making 3D Visual Research Outcomes Transparent* held at the British Academy and the subsequent expert seminar in King's College London in 2006, out of which the LC was born. These two paths started converging when projects started developing more practice-oriented solutions to exemplify how the principles described in the LC (and later the Seville Charter) could be implemented.

2.1 Computational Approaches to Paradata

Already in the mid-1990s, scholars implemented technological solutions to demonstrate uncertainty and make clearer the hypothetical nature of three-dimensional representations. For example, Roberts and Ryan (1997) developed a VRML prototype of a parametric visualisation for the Roman theatre in Canterbury in which the change of different parameters resulted in a change in the 3D representation, thus allowing users to be cognizant of the ambiguity of the preserved evidence. Also, several scholars followed a stylistic approach to highlighting uncertainty. For example, Eiteljorg (2000) used visual degradation and different levels of transparency, while Zuk et al. (2005) implemented various visual cues to highlight temporal uncertainty. Similarly, Roussou and Drettakis (2003) suggested the deployment of non-photorealism for rendering 3D representations in more artistic and expressive styles. Other scholars developed approaches based on computer science and mathematics, for example by suggesting a probabilistic model based on fuzzy logic in which the reliability of 3D representation was given a numerical evaluation (Niccolucci & Hermon, 2010) or a pseudocolour (Sifniotis et al., 2006). A significant amount of 3D scholarship also focused on high fidelity and predictive rendering (see, e.g., Devlin et al., 2003; Happa et al., 2012; Papadopoulos & Earl, 2014) to produce representations that are validated by the simulation of physical properties (e.g. light). Lastly, several projects dealt with uncertainty by developing alternative reconstructions either by means of manual or procedural modelling (Papadopoulos & Earl, 2009; Piccoli, 2016).[2]

2.2 Conceptual Approaches to Paradata: Charters and Principles

The conceptual approach to paradata was born in 2006. It is not entirely clear when the term was first used; however, there is consensus that it was Drew Baker from the King's Visualisation Lab who introduced it in relation to heritage visualisation before it became almost synonymous to the LC. In his paper *Towards Transparency in Visualisation Based Research*, Baker (2007) argued that there is a 'parallel stream of ancillary information to metadata' and that:

> is essential to understanding and building successful and transparent research hypotheses and conclusions, particularly in areas where data is questionable, incomplete or conflicting, and explores how this can be applied to the process of creating three dimensional computer visualisation for research.

In the same paper, and its expanded version in which the term paradata features in the title (Baker, 2012), the author uses the 'Data, Information, Knowledge, Wisdom model' from the field of information science to argue that understanding increases when more connections are created among data and sources, interpretations, and decisions, essentially moving from data—being the lowest level in the knowledge chain—to information, knowledge, and ultimately wisdom. At the same time, however, Baker (2012, 171) criticises the model, by arguing that in the process of metamorphosising data to wisdom, the knowledge chain may get contaminated by both data and processes involved in the move to higher levels of understanding. For this reason, paradata become essential, even though they pose certain challenges

[2] For a more detailed account of uncertainty approaches to 3D modelling in archaeology, see Schäfer (2019).

particularly in terms of quality, quantity, granularity, time, and the sanitisation of the creative process.

The term paradata has become almost synonymous to the LC which was conceived in 2006 as a result of an expert seminar at King's College London organised in the context of the project *Making Space* (run by the King's Visualisation Lab at King's College London and funded by the UK's Arts & Humanities Research Council) in collaboration with the VAST-Lab and EPOCH: The European Network of Excellence on ICT Applications to Cultural Heritage. The first version of the LC entitled 'The London Charter for the Use of Three-dimensional Visualisation in the Research and Communication of Cultural Heritage' was published on the 14th of June 2006, while the second version, entitled 'The London Charter for the Computer-based Visualisation of Cultural Heritage', was published 3 years later, on the 7th February 2009, also accompanied by 'A New Introduction' (Denard, 2012). The second version of the LC aimed at broadening its scope by incorporating not only 3D visualisations but also any other type of computer-based visualisation, including replicas of museum artefacts, as well as those that aim to (re)construct or evoke heritage but do not come from the heritage field, e.g. those in entertainment.

Both versions of the LC described its premises by developing a series of principles rather than concrete aims and methodologies for its implementation. Overall, the LC attempted to create a framework and propose the principles under which visualisation practitioners should operate to achieve intellectual rigour. It is out of the scope of this chapter to delve into the various LC principles; however, it is worth highlighting that although the first version was more open to the selection of the most appropriate means of communication according to the intended message, audience, and circumstances, the second version used much firmer language, and recommendations became granular requirements.[3] For example, Principle 4.5 (2006) suggests that 'it *may be necessary* to disseminate documentation of the interpretative decisions made in the course of a 3d visualisation process and, *as far as is practicable*, the sources used' [emphasis added]; Principle 4.6 (2006) adds that 'transparency information *requirements may change* as levels and sophistication of understanding of particular 3D visualisation methods rise, and *will vary from community to community*' [emphasis added]. In contrast to the above, Principle 4.5 (2009) suggests that 'A *complete list of research sources* used and their provenance should be disseminated' [emphasis added] and Principle 4.6 (2009) adds that 'Documentation of the evaluative, analytical, deductive, interpretative and creative decisions made in the course of computer-based visualisation should be disseminated in such a way that the relationship between research sources, implicit knowledge, explicit reasoning, and visualisation-based outcomes can be understood.'

[3] Even though Denard (2012, 66) comments about the second version that the degree of documentation may depend on the quality of sources and the degree in which a visualisation supports or becomes an argument, this is not evident in the LC principles.

The LC gave birth to other charters and attempts to create frameworks that will implement its principles and standardise paradata documentation. The Seville Charter (SC), for example, was conceived in the context of *The Spanish Society of Virtual Archaeology* as well as the *International Forum of Virtual Archaeology* to develop an implementation guide particularly in relation to issues faced in archaeological 3D visualisation (Grande & Lopez-Menchero, 2011; Lopez-Menchero & Grande, 2011). The beginnings of the SC can be traced back to the establishment of *ARQUEOLÓGICA 2.0: The International Meeting of Archaeology and Graphic Informatics Heritage and Innovation* in 2009, and the session 'Reflections about the London Charter' followed by a plenary assembly on the 'Foundations of Virtual Archaeology'. The first draft of the SC was presented a year later at the second meeting of ARQUEOLÓGICA 2.0. The SC consists of a series of principles some of which overlap with those of the LC. For example, they highlight the need for interdisciplinary collaboration (Principle 1) as well as the use of digital technologies to complement and not replace existing tools and methodologies (Principle 3). The SC ultimately emphasises authenticity as a 'permanent operational concept' (Principle 4) according to which alternative interpretations and different levels of accuracy are presented.[4] This can be achieved through historical rigour (Principle 5), scientific transparency (Principle 7), and the provision of training that will generate more professionals able to conduct and evaluate such scholarship (Principle 8). Although the SC is presented as a means to implement the LC in the context of archaeological heritage, it does not provide specific guidelines or ways to standardise the application of its principles.

2.3 Implementations and Extensions of Paradata Charters and Principles

Several projects have tried to showcase how the London and Seville Charters can be applied to specific case studies. Georgiou and Hermon (2011), for example, use the 3D visualisation of the Hellenistic-Roman theatre in Paphos to provide a list of research sources (Principle 3) and propose ways that Principles 4, 5, and 6 could be implemented, e.g. by developing a reliability chart, applying an XML schema for describing the metadata of the sources that informed the (re)construction, and superimposing the 3D model over the actual remains of the theatre. Hermon and Niccolucci (2018), on the other hand, used the case of the Christ Antiphonitis Church in Kyrenia, Cyprus, to discuss its digital documentation and the virtual recomposition of its frescoes according to the LC principles. However, as the authors admit in their conclusion, objects have features that the principles of the LC cannot capture, also arguing that scientific analyses 'are no less deceptive than a pretty, but

[4] Also see the ICOMOS Charter for the Interpretation and Presentation of Cultural Heritage Sites (Silberman, 2008).

undocumented visualization' (p. 45). Similar projects have also been produced in relation to the SC (see, e.g., Almagro Vidal et al., 2011).

Apart from projects that explicitly follow the London and Seville principles, there are also those that have attempted to extend them by developing new methodological frameworks. For example, Pletinckx (2007) developed the Interpretation Management tool that provides a guide consisting of five steps which mostly focus on LC's Principle 3: research sources, aiming at documenting the correlations among sources and the resulted 3D model to achieve 'scholarly transparency' (p. 5). Using the example of the Saint Saviour church in Ename, Belgium, the author explains the steps that need to be followed for a transparent 3D visualisation (pp. 27–32); however, documenting and linking each source in the proposed manner is not a minor undertaking.

Carrillo Gea et al. (2013) used the LC and the SC as a basis to propose a model for requirements engineering in software development for digital archaeology, while Apollonio and Giovannini (2015), using as a case study the Porta Aurea in Ravenna, developed a rather complex paradata documentation workflow to standardise the capturing of the modelling process and the sources that informed the (re)construction. Grellert et al. (2019) developed 'Sciedoc: The Reconstruction Argumentation Method', a web-based tool for the documentation of decisions in the form of interrelationships among (re)constructions, arguments, and sources. Demetrescu and Ferdani (2021), on the other hand, developed the 'Extended Matrix' by using the underlying principles of the Harris Matrix (Harris, 1989), which was invented in 1973 to describe the physical and temporal relationships between stratigraphic units in archaeological excavations. The Extended Matrix is based on the principle of standardisation by making use of a graph database and a five-step protocol: data collection, analysis, reconstruction, representation, and publication to describe the relationship between the archaeological evidence and the 3D (re)construction. The complexity of the system and the resources required for its implementation may be the reasons why this paradigm has not been embraced by the 3D visualisation community. Lastly, several projects have been developing over the years metadata models and ontologies for the documentation of 3D heritage (see, e.g., Kuroczynski, 2017; for an overview, see Börjesson et al., 2021); however, since their focus is more on standards and sustainability than paradata, their examination is beyond the scope of this chapter.

2.4 Conceptual and Computational Paradata: Looking Back and Looking Forward

This section discussed two approaches to paradata, the conceptual, the beginning of which coincides with the conception of the LC, and the computational, which started more than a decade before the development of the LC. Computational approaches to transparency and uncertainty were developed because of the need to rethink how data uncertainty is communicated to end-users; and, although one could argue that such computational approaches are not a perfect fit for the concept of paradata since they do not document context and decisions, their approach is still based on the fundamental paradata premise, i.e. to communicate intentions, hypotheses, decisions, and interpretations (Börjesson et al., 2021, p. 195). Even though this way of doing paradata may not be as explicit as, for example, a paragraph that describes variables and decisions or data entered in fields with controlled vocabularies, and may not provide the space to develop an argument within the knowledge space of a 3D model (Hoppe, 2001; Schreibman & Papadopoulos, 2019), such computational approaches fulfil the intention to communicate context and decisions at a (visual or other) level that the researcher thought would be appropriate for their audience.

Even though the LC was seen as the Messiah that came to solve the issues that 3D heritage visualisation had faced until then, and undoubtedly shaped later discussions in the field, it has not helped the field to progress. Several projects were developed on the premises of the LC; however, to make it implementable, they followed the path of standardisation, attempting to turn this into another Dublin Core or CIDOC-CRM data model. However, as Sample (2011) argues, paradata is '. . . so flawed, so imperfect that it actually tells us more than compliant, well-structured metadata does.' The downside of the LC and its extensions or implementations is that its principles are based on a seemingly linear process of visualisation in which direct links can be made between data and 3D models. As this chapter demonstrates, tracing roots, tracks, and connections in knowledge production is utterly problematic, especially since various perceptual, physiological, and technical factors, as well as their connections and outcomes, can be rarely identified in a 3D visualisation. The value of approaching knowledge production by documenting research sources in the form of lists and direct relations is debatable, particularly because such documentation would only assert an ostensibly objective method and consequently testify the epistemic authority of the creator.

The next section will lay the foundation for arguing that the concept of paradata in heritage visualisation needs to be revisited. It will do this by problematising the process of 3D (re)construction in archaeology, particularly considering the different aspects of excavating, perceiving, and documenting three-dimensionality.

3 Three-Dimensionality and Knowledge Production in Archaeology

This chapter considers that 3D visualisation is a research project carried out by a hybrid scholar, e.g. a digital archaeologist, who has both domain expertise and highly specialised technical skills. Such a person will produce the 3D visualisation based on their study of relevant materials, including those that may have been produced during an excavation, bibliographic and ethnographic research, the examination of material produced in other research projects, and even discussions with archaeologists and other specialists. Considering that in a 3D visualisation project, one of the first materials we examine are those produced during an excavation, the following sections will explore the perception and documentation of three-dimensionality in relation to excavation practice.

3.1 Excavating the Paradoxes of Archaeology

Archaeology is almost synonymous with excavation (Tilley, 1989, 275); it is a transformative process with ambiguous and paradoxical meaning (Lucas, 2001) that influences the way we understand the past (Edgeworth, 2011). On the one hand, excavation means the recovery of past remains, whereas on the other, to understand and materialise these remains, their context and coherence is shattered, and a new material reality is produced; recording techniques are used to immortalise excavation phases (Bateman, 2006, 68), i.e. to produce a record,[5] which will allow further examination of the evidence. Archaeology is also based on another paradox. Although the real world is three-dimensional, conventional recording mechanisms flatten archaeological reality into a two-dimensional production. Therefore, 3D visualisation specialists rely on the products of this paradox to produce three-dimensional understandings of the past. This transformation of evidence from one form to another is problematic, since we lack the tools, knowledge, and even awareness, to understand the transformations that lead to both a loss and inflow of data. Therefore, in order to critically assess this paradox, we should problematise the various neglected factors that affect the recorded evidence and invalidate the utopian term 'objective record'.

3.2 Perceiving Three-Dimensionality

All objects have a certain morphology; however, to understand the morphology of an object we need to examine both the components that comprise morphology and any contextual elements that influence how these are structured and perceived. More specifically, objects' morphology should be considered both at a micro level, i.e. the fine structure (e.g. colour and texture), and at a macro level, i.e. the gross structure (i.e. geometry and shape). In addition, the processing and construction of information about the real world is based on the principles of three-dimensional vision in coordination with the rest of our sensorium, our situated activities, and embodied practices (Thomas, 1993; Tilley, 1994; Tilley & Bennett, 2004, 2008). However, morphology should also be considered along with contextual or external elements that further define it, such as the light and the angle of view.

Objects and spaces in the world are three-dimensional. However, their optical image formed upon our retina is two-dimensional. This means that our visual system is responsible for transforming this flat image into a three-dimensional representation by using a series of monocular (perceived by the operation of one eye) and binocular (perceived by the operation of both eyes) cues. We can see in three dimensions because of our retinal disparity, i.e. the use of our eyes located at a different position in our head (stereopsis), which provides the information needed

[5] For a critique of the notion of the 'record', see Barrett (1988, 2006); Edgeworth (2003); Hamilakis (1999); Patrick (1985).

by the brain to calculate depth. Convergence and accommodation (Helmholtz, 1856) are used together with stereopsis to focus a scene on the retina. When movement is involved, motion parallax, a movement-produced cue related to the motion of the observer, facilitates depth perception also leading to the accretion and deletion of objects as we move relative to them (Gibson et al., 1959; Rogers & Graham, 1979).

The perception of distance and object size also depends on a series of monocular cues. For example, perspective is one of the most known and well-understood cues, as it is based on a simple principle; the object that is closer to the eyes appears larger, whereas the further the object, the smaller its retinal image will be. Based on the principles of perspective there are also other cues that influence depth perception; these include linear (Saunders & Backus, 2006) and atmospheric perspective (O'Shea et al., 1994), texture gradient (Gibson, 1950, 77–94), relative and familiar size (Hochberg & Hochberg, 1952; Hochberg & McAlister, 1955), relative height (Dunn et al., 1965; Epstein, 1966), and interposition/occlusion (Chapanis & McCleary, 1953). However, some sources of depth perception provide more reliable signs of depth than others (Guibal & Dresp, 2002; Hillis et al., 2004; Jacobs, 2002).

Our sight functions in coordination with light. We can see because particles of light bounce on objects and surfaces, then reach our eyes, and in turn this information is deciphered in the brain (Tarr et al., 1998; Wade, 1999, 9–25). The initial processing of light patterns takes place at the retina of our eye which is layered with cone photoreceptors, sensitive to wavelengths of red (R), green (G), and blue (B) colour (Kaiser & Boynton, 1996). Light, in a combination with our existing knowledge of the world, also generates two phenomena that enable the perception of objects' morphology and spatial relationships: (1) shading, i.e. the variation of light's intensity on different surfaces, generated by light coming from a particular angle and reflected off surfaces in a particular way (Kleffner & Ramachandran, 1992; Ramachandran, 1988), and (2) shadowing, i.e. the way that shadows are cast when an object blocks the path of light onto another one (Cavanagh & Leclerc, 1989; Mamassian et al., 1998). Although our brain uses shadows and shading to extract information to enhance depth perception, research suggests (Ho et al., 2006, 645–646) that observers can make errors regarding surfaces' roughness.

The position of the sun, the clouds in the sky, and the haze of the atmosphere make light behave in different ways and consequently affect the perception of objects illuminated under these changing conditions. Objects' reflectance and transmittance properties also affect light's behaviour. Furthermore, the three variables of light, i.e. intensity, distribution, and color, greatly vary depending on the source of illumination, thus further affecting how object morphology is perceived. Ashley (2008) undertook systematic vision testing under different lighting conditions in the excavation of Çatalhöyük, demonstrating that people perceive environments depending on several internal and external dynamics, therefore making evident the need for a viewer-centred archaeology.

The perception of three-dimensional space is a multimodal production since objects stimulate all the sensory organs of the human body. This process, however, is not linear since sensory systems are triggered differently depending on objects' properties. For example, we have learnt to associate texture with tactility (Klatzky & Lederman, 1987; Lederman & Klatzky, 1987; Taylor et al., 1973); however, the initial information about objects' texture is extracted from the visual system, which then directs the other perceptual mechanisms to enhance objects' surface perception (Heller, 1982; Landy & Graham, 2004). In addition, situated activities and embodied practices, experiences (Charest, 2009), memories (Casey, 2000; Jones, 2007, 1–26), and the emotional and motivational state of the observer significantly affect the way that reality is perceived. In other words, three-dimensional space is an amalgamation of visual learning and intuition (Gibson, 1950, 10–16).

3.3 Flattening the Three-Dimensional World into Two-Dimensional Records

Conventional recording methods, such as text, drawing, and photography, depict three-dimensionality with a series of conventions based on established and to a great extent blackboxed practices (Latour, 1999; Lucas, 2012, 239). These attempt to separate the subjective from the objective (Barker, 1993, 163; Yarrow, 2003, 72) and to ensure that any biases can be identified (Andrews et al., 2000, 526). Although this chapter takes as a premise the normative mode of translation in archaeological practice, i.e. from three to two dimensions, it does not assume that this is always the case. An increasing number of projects deploy 3D methods of documentation, while others (see, e.g., Dawson and Reilly, chapter 'Towards Embodied Paradata. A Diffractive Art/Archaeology Approach' in this volume; Reilly et al., 2021) have been exploring multimodal translations, e.g. sound, to problematise the nature of the archaeological record. This chapter, however, does not aim at addressing these separately since it is believed that regardless of the medium used for capturing information, the factors that affect processes and outcomes, including technological authority, individual choices, and sensory perception, overlap.

Three-Dimensionality in Text

The most common recording method in archaeological practice is text (Hodder, 1989). The objective—subjective polarity of processualists and post-processualists—gave birth to different methods for recording an excavation in textual form. Since it was thought that written records, and especially these in the form of descriptive narrative, cannot express the excavation as a neutral and scientific record, Single Context Recording (Westman, 1994, §1.2) and Harris Matrices (Harris, 1989) were employed, partially replacing discursive field diaries. The predefined forms and detailed guidance that these provided attempted to ensure that results retain their neutrality regardless of the agents of excavation and their actions (Edgeworth, 2003).

Textual sources can provide a wide range of information regarding the perception of three-dimensionality; for example, in notebooks, where descriptive narrative is mainly used, the identification of colour and texture depends on an individual's observation and free description. Inventions such as the Munsell Color Chart

(Munsell, 1905, 1912) provide some standardisation, but many conditions must be met; for example, the readings should be taken under natural light, on a sunny day, and the soil should be moist. Such parameters confirm that 'The probability of having a perfect matching is less than one in one hundred' (Munsell Soil Color Charts, 1994, 1). Goodwin (2000) examined the process of defining the colour of soil in excavations suggesting that it is not only a mental task but also a situated activity which involves physical tools and embodied practices, and thus people perceive and describe colours differently. Similar problems arise when describing texture. Although there are flowcharts that help in the identification of the texture of soils and sediments by finger testing, this is also a subjective process determined by the individuals who record the evidence.

Three-Dimensionality in Photography

Photography was adopted by archaeology soon after its invention, as it was believed that in that way any subjectivity could be overcome by becoming the memories of the past transformed in the excavation (Locatelli et al., 2011, 329). A number of factors invalidate the claim that photography produces an objective pictorial record compared to other illustrative methods (Conlon, 1973, 55; Ivins 1953, 137). For example, technical parameters, such as lens quality, the format, and processing. affect how reality is captured. Cameras are also inherently limited in distinguishing subtle colour/tone changes, while poor exposure latitude, i.e. the range between the lightest and darkest parts, should also be considered (Hester et al., 1997, 166). Colour capturing also depends on the type and sensitivity of sensors and also varies depending on the reproduction medium, e.g. a computer monitor or a printer. The relative position of the photographer, the angle of view, and the distance from the subject also have an impact on the understanding of a captured scene.

Photographs do not objectively capture, but they possess an interpretive role which derives from the different kinds of gaze ingrained in the photograph and accrued from its context (Lutz & Collins, 2003). In archaeology, photographs are used out of context, along with other images and text focusing on specific aspects; therefore, they are to a certain extent manipulated to represent in a seemingly unbiased manner a particular moment in time. The pluritemporality of the sites is therefore lost (Dawson et al., 2022).

Three-Dimensionality in Drawing

Drawing in archaeology is still synonymous to pen and paper, helping archaeologists to decipher material relationships which are not understandable by any other means. Schematic, interpretative, or pictorial/naturalistic and highly conventionalised drawings transform a three-dimensional, colourful, and freely defined real world into a flat, linear, and colourless production (Leibhammer, 2000, 129; Piggott, 1965, 165). Excavated features are translated into flat lines: edges become fixed, silhouettes clearly defined, and black lines delineate space (Ford, 1993, 319). In the physical world, however, objects are not flat and do not have clear edges, while outlines are diffuse and multiple. Therefore, drawing diminishes the sense of three-dimensionality, while personal choices, the angle of view, and perspective distortion cause further misjudgements regarding shapes and edges (Griffith et al., 1996, 97). Also, colour variation in soil, which is essential for understanding slight changes in contexts, is not depicted in drawings, which are typically in black and white format. The depiction of texture by using stippling, hatches, lines, and gradations of tone is equally problematic, as it relies on project-specific conventions, and in most cases, little indication of texture is included in the drawings or in field notes. Guides for good practice also suggest that light and shade should be omitted; otherwise drawings may be misty and confusing (Griffith et al., 1996, 100).

Drawings are subjective responses to the immaterialisation of the world, and as such, they always vary; this is not only due to different perceptual capacities and skills but also due to illustrators' style and viewpoint and their decisions about what to include and omit (Morgan & Wright, 2018). Illustration, as is the case with photography and text, is an interpretative act.

4 A Leap of Faith: Revisiting Paradata

Archaeological remains are translated into different chronotopes, both during an excavation and during documentation, study, and (re)construction. Although we argue for the need of more and better provenance documentation (Reilly et al., 2021), the identification of provenance becomes a complex multifaceted pursuit since the origins of the decisions we make and of the materials produced are framed and afforded by data structures and standards, conventions, limitations of tools and methods, cognitive mechanisms, and personal capacities. As a result, the argument that we can go back to the initial information or that the translation process can be circulated (Witmore, 2004) should be challenged.

Many scholars have addressed the seer complexity of paradata, especially in the case of (3D) heritage visualisation. Turner (2012), for example, argues that the formation of understandings via visual perception can be complex but also confusing and wonders if paradata could provide a solution or a curse. Devlin (2012) also addresses the complex nature of computer graphic simulation arguing that there are many factors, including the inherent limitations of technology as well as visual perception that make transparency challenging. Reilly et al. (2021), on the other hand, by applying an art/archaeology approach to archaeological practice, discuss the ontological shifts that conventional recording methods undergo and argue that paradata become elusive and illusionary. Similarly, Börjesson et al. (2022) highlight paradata's technical and epistemological heterogeneity and the challenges in identifying and analysing them due to their different levels of completion, writing style, and nomenclature. Other scholars have also gone a step further, suggesting that the concept may be counterintuitive and may need to be reconsidered or even abandoned. For example, Havemann (2012) writes about naive paradata (p. 158) and proposes a more 'reasonable' approach (p. 159) in which only meaningful paradata are preserved. Mudge (2012) suggests that paradata may have to be retired and replaced by the Lab Notebook, while Schreibman and Papadopoulos (2019) argue that even if it was possible to document every decision, documenting and representing the rationale for such decisions is an unobtainable task.

The concept of paradata, which has become almost synonymous with the London Charter, has haunted the field of (3D) heritage visualisation. This is because, although post-processual in nature—aiming at making space for variables and multiple interpretations—its elaborations and suggested implementations have been largely underlined by a processual discourse according to which rigour and scientificity can make processes reproducible. In 2006, when the term was first introduced in 3D heritage visualisation, the field was carrying a heavy baggage: that of photorealism and constant shifts in technology. However, 15 years later, and considering both the evident lack of systematic integration of paradata practices in 3D visualisation, as well as the move towards alternative forms of scholarship, the recognition of atypical outputs, and the renegotiation of established norms and practices, paradata need to be reconsidered. This chapter argues that although paradata is a very much needed concept, it requires a leap of faith.

Attempts to standardise the documentation of paradata and take away the roughness inherent in the processes and protocols employed in archaeological practice (Börjesson et al., 2022) do not do justice to the richness and flexibility that the concept of paradata provides. The need for standardisation has been dictated by the emphasis that the research community has placed on issues of transparency, without considering that the inherent problems are not dissimilar to the issues faced in conventional means of representation and research outputs. Two decades ago, scholars started embarking on photorealism, and therefore the need for transparency and authenticity emerged. This was also deemed to be the means through which the 3D scholarly community could respond to the criticisms of a more conservative research community that did not have the capacity to deal with outputs that were not part of the canon. However, we have had enough exposure to such products over the years, and thus we are more able to evaluate 3D scholarship. Although we may still require new literacies to decipher its products, it is not reasonable to put the entire burden on the creators of such outputs, by asking them to transform their scholarship into forms that correspond more closely to the research outputs our field has been accustomed to.

How reasonable would it be to ask historians, who, for example, write about a historical event, to compile lists with sources, correlations, and hypotheses in order to prove that there is a linear relationship between the sources and their interpretation? If this is not an expectation we have from a historian or any other humanities scholar, why should a 3D visualisation scholar be an exception? Why should a 3D visualisation be accompanied by additional documentation that accounts for every decision and the factors that influenced those? We need to accept that scholars who use 3D for analysis, synthesis, and knowledge communication have the necessary scholarly expertise to decide what aspects of their decision-making need to be communicated to their intended audience. Consequently, we need to trust that the recipients of that scholarship have a sufficient understanding of what such scholarship entails and, thus, can evaluate research outputs. At the same time, we need to accept that our processes have inherent biases and contaminations. As

Baker (2007), who first used the term paradata in heritage visualisation, argued, in the process of creating connections to transform data to wisdom, the knowledge chain will get contaminated by conscious and unconscious variables.

How to do paradata then? Is there a minimum level of detail and an appropriate form for our information gathering and decision-making that will be adequate? And is it only paradata that we need to communicate or also the peridata (Gant & Reilly, 2018), i.e. the decisions about what has been included or omitted as paradata? Is this even feasible or useful? And how can or should we account for all the chronotopes and pluritemporalities (Dawson et al., 2022; Reilly et al., 2021) we produce in our practice and the data hidden in our conscious and unconscious processes? Do we need to move towards the systematisation of paradata or could we see the inherent roughness in our practice as an opportunity for reflection and self-expression (Börjesson et al., 2022)? Lastly, since changes in technology and obsolescence seem to be inevitable in 3D scholarship, should paradata also be able to dynamically adapt to the changing ecosystems we work in? This chapter does not offer a response to any of these questions but only a perspective that problematises the great range of variables that influence knowledge production, thus demonstrating that the guidelines set by such charters must be revisited. Tracing the interpretative process through lists or trees of hypotheses becomes onerous and counterintuitive and neglects that decision-making is sensory, embodied, and multitemporal, as well as a sociocultural, situated act.

3D modelling, similar to recording in excavation, is not a passive transformative process but a choreographed (Huvila & Sköld, 2021) worlding (Pijpers, 2021), which makes the modeller think about the translation of the archaeological material into a computer programme. Such programmes enable their operators, through tools and conventions to produce the attributes of three-dimensional space; to do that, they also require skills, and thus, personal capacities and choices, while the affordances of technology also play a major role in this process. Therefore, similarly to the process of documentation during an excavation that is dependent on a wide range of often neglected perceptual, physiological, and technical factors—therefore generating an impenetrable black box—mechanisms of reproduction should also be challenged, not necessarily to dismantle their black boxes but at least to raise awareness of the variables and factors that invalidate the argument that 3D visualisation should be a reproducible act.

The author has argued together with Schreibman (Papadopoulos & Schreibman, 2019; Schreibman & Papadopoulos, 2019) that there is an imperative need to move towards a different paradigm and has proposed the theoretical and methodological framework of 3D Scholarly Editions: a framework that allows the production of an ecosystem around 3D scholarship that has the potential to enable and stimulate the scrutiny of authenticity and the rethinking of what paradata should be and how should be captured. By looking at 3D as a form of text, we are permitted to build an intertextual network that provides the potential for linking the editorial, epistemological, and technical processes involved in 3D knowledge production. Thinking of 3D as text is not problem-free and adds further complexity to an already complex process of interpretation. For example, who is the author of that text and what is the role of the editor? Is the 3D modeller the author and is the editor the person who annotates and contextualises the model? In this paradigm we have argued that the goal of a 3D Scholarly Edition is not to remediate the intention of the author (i.e. the modeller) but that the modeller is another kind of editor in the text's (re)construction. There are also further complications to this model, especially if we think that the role of the editor can also be assumed by non-human actors, e.g. in the case of dynamic annotations and Linked Open Data. In this model, we do not propose to see the editor as someone who testifies epistemic authority in the process of knowledge production, but as someone who is allowed (and enabled by a 3D Scholarly Edition technological framework) to construct a knowledge site that will provide the scholarly community with tools for 'prying problems apart and opening up a new space for the extension of learning' (Apollon et al., 2014, 5–6).

The leap of faith is presented here not just as a colloquial concept but also as a framework that opens up new possibilities for looking at paradata—especially in the context of assessment reform and non-typical outputs—and breaking away from the originally suggested rigidity and standardisation. While the connection between paradata and research assessment may not seem obvious, it is important to consider that paradata in heritage visualisation was suggested as a means to promote transparency, and in response to the criticism that 3D scholarship failed to adhere to established standards and practices. In this regard, the leap of faith provides a conceptual framework for ensuring that research processes and outputs are open, transparent, and inclusive; it emphasises the importance of diverging conventional notions of scholarship while also trusting researchers' ethics and integrity. This is in line with the recent Declaration on Research Assessment (DORA, n.d.), the Agreement on Reforming Research Assessment (CoARA, 2022), and the Dutch Recognition and Rewards programme (R&R, n.d.), as well as discussions, especially in Digital Humanities (see, e.g., Nyhan, 2020; Schreibman et al., 2011), about expanding the understanding of what scholarship means and how to recognise and evaluate work that falls outside of conventional venues. Seeing paradata through a leap of faith, then, can facilitate this process and smoothen the transition to more contemporary conceptions of scholarship.

5 Conclusion

Using as a starting point the principles set out by the LC and its various implementations, this chapter attempted to look back at conventional archaeological practice and problematise processes and products of interpretation. The premise of this chapter is that the creation of a 3D (re)construction requires us to look back at the unearthing of data and try to decipher the processes deployed in their documentation. By presenting the principles of three-dimensionality, both in terms of perception and recording, this chapter showcases that a 3D (re)construction is not a linear process and does not happen in a single black box since every element of the visualisation process is by its nature bounded in a black box. Since knowledge is built through perception, and individuals' perceptual abilities vary, the mechanisms of knowledge production, and consequently the resulted knowledge, vary too. Perception is a complex mechanism, influenced not only by our senses but also by our experiences and memories. Besides, it is our body, which is the decisive factor in the formation of understandings, by providing the sensoria through which experiences about the world are structured.

Considering that there are such complex processes that make 'the joint production of actors and artifacts entirely opaque' (Latour, 1999, 183), this chapter proposes that the concept of paradata—at least in how it has been interpreted by the LC, needs to be revisited. Instead of arguing for a process that will standardise the capturing of paradata, e.g., to make them machine readable, 3D visualisation requires a new approach—what I call a leap of faith—that aligns with our increased capacity to deal with, as well as recognise, evaluate, and reward 3D scholarship. The inherent roughness and lack of systematicity that 3D visualisation entails and the fact that paradata is 'bound to be incomplete' (Huvila, 2022) should be seen as an opportunity to develop new frameworks that will enable the authors and editors of 3D models to break free from the shackles of the LC and develop embodied productions of materiality that can do justice to the 'unbearably complex' (Huvila, 2013) nature of 3D (re)construction. In such a way, the reconceptualisation of paradata within a framework that allows us to produce 3D scholarship that can be seen as equal to other forms of (digital) scholarship can provide the means to better integrate less typical outputs into our fields and thus expand the textual, visual, and multimodal vocabularies of knowledge production.

Acknowledgements The author would like to thank Paul Reilly for his feedback on an earlier iteration of this chapter as well as the editors and the anonymous reviewers for their valuable comments and suggestions.

This chapter is published in the context of the project PURE3D: An Infrastructure for the Publication and Preservation of 3D Scholarship, funded by the Platform Digitale Infrastructuur—Social Sciences and Humanities (PDI-SSH) in the Netherlands.

References

Almagro Vidal, A., Gómez Merino, J. L., & Ramírez González, R. (2011). The Toledo Gate in Ciudad Real, Spain. An applied case study of the Seville charter. In K. Pavelka (Ed.), *Proceedings of XXIIIrd international CIPA symposium, Prague, Czech Republic, 12–16 September 2011*. Retrieved September 17, 2022, from https://www.cipaheritagedocumentation.org/activities/conferences/proceedings_2011/.

Andrews, G., Barrett, J. C., & Lewis, J. S. C. (2000). Interpretation not record. The practice of archaeology. *Antiquity, 74*, 525–530. https://doi.org/10.1017/S0003598X00059871

Apollon, D., Bélisle, C., & Régnier, P. (2014). *Digital critical editions*. University of Illinois Press.

Apollonio, F. I., & Giovannini, E. C. (2015). A paradata documentation methodology for the Uncertainty Visualization in digital reconstruction of CH artifacts. *SCIRES-IT-SCIentific RESearch and Information Technology, 5*(1), 1–24. https://doi.org/10.2423/i22394303v5n1p1

Ashley, M. (2008). Towards an archaeology of vision: Observations from the site of Çatalhoyuk, Turkey. In J. Thomas & V. Oliveira-Jorge (Eds.), *Archaeology and the politics of vision in a post-modern context* (pp. 101–117). Cambridge Scholars Publishing.

Baker, D. (2007, June 19). Towards transparency in visualisation based research. In *Proceedings of the seminar from abstract data mapping to 3D photorealism: Understanding emerging intersections in visualisation practices and techniques*. Birmingham Institute of Art and Design. Retrieved September 19, 2022, from http://www.kvl.cch.kcl.ac.uk/makingspace/ttivbr.html.

Baker, D. (2012). Defining paradata in heritage visualization. In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.), *Paradata and transparency in virtual heritage* (pp. 163–175). Routledge.

Barker, P. (1993). *Techniques of archaeological excavation* (1st ed. 1977) (3rd ed.). Batsford.

Barrett, J. C. (1988). Field of discourse: Reconstituting a social archaeology. *Critique of Anthropology, 7*, 5–16. https://doi.org/10.1177/0308275X88007003

Barrett, J. C. (2006). Archaeology as the investigation of the contexts of humanity. In D. Papaconstantinou (Ed.), *Deconstructing context. A critical approach to archaeological practice* (pp. 194–211). Oxbow Books.

Bateman, J. (2006). Pictures, ideas, and things: The production and currency of archaeological practice. In M. Edgeworth (Ed.), *Ethnographies of archaeological practice: Cultural encounters material transformations* (pp. 68–80). Altamira Press.

Beacham, R. (2012). Defining our terms in heritage visualization. In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.), *Paradata and transparency in virtual heritage* (pp. 33–38). Routledge.

Beacham, R., Denard, H., & Niccolucci, F. (2006). An introduction to the London Charter. In M. Ioannides et al. (Eds.), *The e-volution of information communication and technology in cultural heritage, proceedings of VAST 2006* (pp. 263–269). Archaeolingua.

Biederman, I. (2001). Recognizing depth-rotated objects: A review of recent research and theory. *Spatial Vision, 13*(2–3), 241–253. https://doi.org/10.1163/156856800741063

Börjesson, L., Sköld, O., Friberg, Z., Löwenborg, D., Pálsson, G., & Huvila, I. (2022). Re-purposing excavation database content as paradata: An explorative analysis of paradata identification challenges and opportunities. *KULA: Knowledge Creation, Dissemination, and Preservation Studies, 6*(3), 1–18. https://doi.org/10.18357/kula.221

Börjesson, L., Sköld, O., & Huvila, I. (2021). Paradata in documentation standards and recommendations for digital archaeological visualisations. *Digital Culture & Society, 6*(2), 191–220. https://doi.org/10.14361/dcs-2020-0210

Callieri, M., Dell'Unto, N., Dininno, D., & Ekengren, F. (2020). *Dynamic collections: A 3D web infrastructure designed to support higher education and research in archaeology*. Retrieved October 2, 2022, from https://www.darklab.lu.se/digital-collections/dynamic-collections/

Carrillo Gea, J. M., Toval, A., Fernández Alemán, J. L., Nicolás, J., & Flores, M. (2013). The London Charter and the Seville Principles as sources of requirements for e-archaeology systems development purposes. *Virtual Archaeology Review, 4*(9), 205–211. https://doi.org/10.4995/var.2013.4275

Casey, E. S. (2000). *Remembering: A phenomenological study* (1st ed. 1987) (2nd ed.). Studies in continental thought. Indiana University Press.

Cavanagh, P., & Leclerc, Y. G. (1989). Shape from shadows. *Journal of Experimental Psychology Human Perception and Performance, 15*, 3–27. https://doi.org/10.1037/0096-1523.15.1.3

Chapanis, A., & McCleary, R. A. (1953). Interposition as a cue for the perception of relative distance. *The Journal of General Psychology, 48*(2), 113–132. https://doi.org/10.1080/00221309.1953.9920186

Charest, M. (2009). Thinking through living: Experience and the production of archaeological knowledge. *Archaeologies: Journal of the World Archaeological Congress, 5*(3), 416–445. https://doi.org/10.1007/s11759-009-9116-x

CoARA. (2022). Coalition for advancing research assessment. Retrieved February 13, 2023, from https://coara.eu/.

Conlon, V. M. (1973). *Camera techniques in archaeology*. John Baker.

Couper, M. P. (2017). Birth and diffusion of the concept of paradata. *Advances in Social Research, 18*, 14–26, in Japanese (Trans. W. Matsumoto). Retrieved September 4, 2022, from http://www.kyoto-info.com/kyoto/books/socialresearch/18.html, English version: http://jasr.or.jp/english/JASR_Birth%20and%20Diffusion%20of%20the%20Concept%20of%20Paradata.pdf.

Dawson, I., Jones, A. M., Minkin, L., & Reilly, P. (2022). Temporal Frankensteins and legacy images. *Digital, 2*(2), 244–266. https://doi.org/10.3390/digital2020015

Demetrescu, E., & Fanini, B. (2017). A white-box framework to oversee archaeological virtual reconstructions in space and time: Methods and tools. *Journal of Archaeological Science: Reports, 14*, 500–514. https://doi.org/10.1016/j.jasrep.2017.06.034

Demetrescu, E., & Ferdani, D. (2021). From field archaeology to virtual reconstruction: a five steps method using the extended matrix. *Applied Sciences, 11*(11), 5206. https://doi.org/10.3390/app11115206

Denard, H. (2012). A new introduction to the London Charter. In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.), *Paradata and transparency in virtual heritage* (pp. 57–71). Routledge.

Devlin, K. (2012). Just how predictable is predictive lighting? In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.), *Paradata and transparency in virtual heritage* (pp. 151–209). Routledge.

Devlin, K., Chalmers, A., & Brown, D., (2003). Predictive lighting and perception in archaeological representations. In *UNESCO World heritage in the digital age. Proceedings of the 30th anniversary digital congress*. UNESCO World Heritage Centre. Retrieved September 17, 2022, from http://doc.gold.ac.uk/~mas01kd/publications/unesco_paper.pdf

DORA. (n.d.). *San Francisco Declaration on Research Assessment*. Retrieved September 30, 2022, from https://sfdora.org/.

Dunn, B. E., Gray, G. C., & Thompson, D. (1965). Relative height on the picture plane and depth perception. *Perceptual and Motor Skills, 21*(1), 227–236. https://doi.org/10.2466/pms.1965.21.1.227

Edgeworth, M. (2003). *Acts of discovery: An ethnography of archaeological practice* (BAR international series 1131). Archaeopress.

Edgeworth, M. (2011). Excavation as a ground of archaeological knowledge. *Archaeological Dialogues, 18*(1), 44–46. https://doi.org/10.1017/S1380203811000109

Eiteljorg, H., II. (2000). The compelling computer image: A double-edged sword. *Internet Archaeology, 8*. Online. https://doi.org/10.11141/ia.8.3

Epstein, W. (1966). Perceived depth as a function of relative height under three background conditions. *Journal of Experimental Psychology, 72*(3), 335–338. https://doi.org/10.1037/h0023630

European Commission. (2021, November 10). *Commission recommendation on a common European data space for cultural heritage.* Retrieved September 18, 2022, from https://digital-strategy.ec.europa.eu/en/news/commission-proposes-common-european-data-space-cultural-heritage.

European Commission. (2022, April 25). *Study on quality in 3D digitisation of tangible cultural heritage: Mapping parameters, formats, standards, benchmarks, methodologies, and guidelines*. Retrieved September 18, 2022, from https://digital-strategy.ec.europa.eu/en/library/study-quality-3d-digitisation-tangible-cultural-heritage.

Ford, D. (1993). The nature of clarity in archaeological line drawings. *Journal of Field Archaeology, 20*(3), 319–333. https://doi.org/10.2307/530056

Frieman, C., & Gillings, M. (2007). Seeing is perceiving? *World Archaeology, 39*(1), 4–16. https://doi.org/10.1080/00438240601133816

Gant, S., & Reilly, P. (2018). Different expressions of the same mode: A recent dialogue between archaeological and contemporary drawing practices. *Journal of Visual Art Practice, 17*(1), 100–120. https://doi.org/10.1080/14702029.2017.1384974

Georgiou, R., & Hermon, S. (2011). A London Charter's visualization: the ancient Hellenistic-Roman theatre in Paphos. In M. Dellepiane, F. Niccolucci, S. Pena Serna, H. Rushmeier, & L. Van Gool (Eds.), *The 12th international symposium on virtual reality, archaeology and cultural heritage VAST* (pp. 53–56). The Eurographics Association). https://doi.org/10.2312/PE/VAST/VAST11S/053-056

Gerardin, P., Kourtzi, Z., & Mamassian, P. (2010). Prior knowledge of illumination for 3D perception in the human brain. *Proceedings of the National Academy of Sciences of the United States of America, 107*(37), 16309–16314. https://doi.org/10.1073/pnas.1006285107

Gibson, E. J., Gibson, J. J., Smith, O., & W., Flock, H. (1959). Motion parallax as a determinant of perceived depth. *Journal of Experimental Psychology, 58*(1), 40–51. https://doi.org/10.1037/h0043883

Gibson, J. J. (1950). *The perception of the visual world*. Houghton Mifflin.

Goodwin, C. (2000). Practices of color classification. *Mind, Culture and Activity, 7*(1-2), 19–36. https://doi.org/10.1080/10749039.2000.9677646

Grande, A., & Lopez-Menchero, V. M. (2011). The implementation of an international charter in the field of virtual archaeology. In K. Pavelka (Ed.), *Proceedings of XXI-IIrd international CIPA symposium, Prague, Czech Republic, 12–16 September 2011*. Retrieved September 17, 2022, from https://www.conferencepartners.cz/cipa/proceedings/pdfs/B-2%20Seville%20charter/Grande%20Leon.pdf.

Grellert, M. A. R. C., Apollonio, F. I., Martens, B., & Nubbaum, N. (2019). Working experiences with the reconstruction argumentation method (RAM)—Scientific documentation for virtual reconstruction. In W. Börner & S. Uhlirz (Eds.), *23rd international conference on cultural heritage and new technologies (CHNT 23)*. Museen der Stadt Wien—Stadtarchäologie. Retrieved October 2, 2022, from https://archiv.chnt.at/wp-content/uploads/eBook_CHNT23_Grellert.pdf

Griffith, N., Jenner, A., & Wilson, C. (1996). *Drawing archaeological finds: A handbook*. Archetype.

Guibal, C., & Dresp, B. (2002). The perception of apparent depth: From cue combination to cue competition. *Perception, 31*(Suppl) European Conference on Visual Perception Abstract Supplement. https://doi.org/10.1177/03010066020310S101

Hamilakis, Y. (1999). La Trahison des Archéologues? Archaeological practice as intellectual activity in postmodernity. *Journal of Mediterranean Archaeology, 12*(1), 60–79. https://doi.org/10.1558/jmea.v12i1.60

Happa, J., Bashford-Rogers, T., Wilkie, A., Artusi, A., Debattista, K., & Chalmers, A. (2012). Cultural heritage predictive rendering. *Computer Graphics Forum, 31*(6), 1823–1836. Blackwell. https://doi.org/10.1111/j.1467-8659.2012.02098.x

Harris, C. E. (1989). *Principles of archaeological stratigraphy* (1st ed 1979) (2nd ed.). Academic Press.

Havemann, S. (2012). Intricacies and potentials of gathering paradata in the 3D modelling workflow. In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.), *Paradata and transparency in virtual heritage* (pp. 220–235). Routledge.

Heller, M. A. (1982). Visual and tactual texture perception: Intersensory cooperation. *Perception & Psychophysics, 31*(4), 339–344. https://doi.org/10.3758/BF03202657

Helmholtz, H. (1856). Ueber die Accommodation des Auges. *Graefe's Archiv für ophthalmologie, 2*, 1–74. https://doi.org/10.1007/BF02720789

Hermon, S., & Niccolucci, F. (2018). Digital authenticity and the London Charter. In P. Di Giuseppantonio Di Franco, F. Galeazzi, & V. Vassallo (Eds.), *Authenticity and cultural heritage in the age of 3D digital reproductions* (pp. 37–47). McDonald Institute for Archaeological Research.

Hester, T. R., Shafer, H. J., & Feder, K. L. (1997). *Field methods in archaeology* (1st edition 1949 by Heizer, R.) (7th ed.). Mayfield.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision, 4*, 967–992. https://doi.org/10.1167/4.12.1

Ho, Y.-X., Landy, M. S., & Maloney, L. T. (2006). How direction of illumination affects visually perceived surface roughness. *Journal of Vision, 6*(5), 634–648. https://doi.org/10.1167/6.5.8

Ho, Y.-X., Maloney, L. T., & Landy, M. S. (2007). The effect of viewpoint on perceived visual roughness. *Journal of Vision, 7*(1), 1–16. https://doi.org/10.1167/7.1.1

Hochberg, C. B., & Hochberg, J. E. (1952). Familiar size and the perception of depth. *The Journal of Psychology, 34*(1), 107–114. https://doi.org/10.1080/00223980.1952.9916110

Hochberg, J. E., & McAlister, E. (1955). Relative size vs. familiar size in the perception of represented depth. *The American Journal of Psychology, 68*(2), 294–296. https://doi.org/10.2307/1418903

Hodder, I. (1989). Writing archaeology: Site reports in context. *Antiquity, 63*, 268–274. https://doi.org/10.1017/S0003598X00075980

Hodder, I. (2000). Developing a reflexive method in archaeology. In I. Hodder (Ed.), *Towards reflexive method in archaeology: the example at Çatalhöyük* (pp. 3–14). McDonald Institute for Archaeological Research.

Hoppe, S. (2001). Die Fußnoten des Modells. CAD-Modelle als interaktive Wissensräume am Beispiel des Altenberger-Dom-Projektes. In M. Frings (Ed.), *Der Modelle Tugend. CAD und die neuen Räume der Kunstgeschichte* (pp. 87–102). Weimar.

Huvila, I. (2013). The unbearable complexity of documenting intellectual processes: Paradata and virtual cultural heritage visualisation. *Human IT: Journal for Information Technology Studies as a Human Science, 12*(1) Retrieved September 2, 2022, from https://humanit.hb.se/article/download/96/82

Huvila, I. (2018). The subtle difference between knowledge and 3D knowledge. *Hamburger Journal für Kulturanthropologie, 7*, 99–111. https://nbn-resolving.org/urn:nbn:de:gbv:18-8-11966

Huvila, I. (2022). Improving the usefulness of research data with better paradata. *Open Information Science, 6*(1), 28–48. https://doi.org/10.1515/opis-2022-0129

Huvila, I., & Sköld, O. (2021). Choreographies of making archaeological data. *Open Archaeology, 7*, 1602–1617. https://doi.org/10.1515/opar-2020-0212

Ivins, W. M., Jr. (1953). *Prints and visual communication*. The MIT Press.

Jacobs, A. R. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences, 6*(8), 345–350. https://doi.org/10.1016/s1364-6613(02)01948-4

Jones, A. (2007). *Memory and material culture* (Topics in contemporary archaeology series). Cambridge University Press.

Kaiser, P. K., & Boynton, R. B. (1996). *Human color vision* (2nd ed.). Optical Society of America.

Klatzky, R. L., & Lederman, S. J. (1987). *The intelligent hand*. Academic Press.

Kleffner, D. A., & Ramachandran, V. S. (1992). On the perception of shape from shading. *Perception & Psychophysics, 52*, 18–36. https://doi.org/10.3758/BF03206757

Kuroczynski, P. (2017). Virtual research environment for digital 3D reconstructions–standards, thresholds and prospects. *Studies in Digital heritage, 1*(2), 456–476. https://doi.org/10.14434/sdh.v1i2.23330

Landy, M. S., & Graham, N. (2004). Visual perception of texture. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1106–1118). MIT Press.

Latour, B. (1999). *Pandora's hope. Essays on the reality of science studies*. Harvard University Press.

Lederman, S. J., & Klatzky, R. L. (1987). Hand movements: A window into haptic object recognition. *Cognitive Psychology, 19*(3), 342–368. https://doi.org/10.1016/0010-0285(87)90008-9

Leibhammer, N. M. (2000). Rendering realities. In I. Hodder (Ed.), *Towards reflexive method in archaeology: the example of Çatalhöyük* (pp. 129–142). McDonald Institute for Archaeological Research.

Locatelli, P. M., Simone, C., & Ardesia, V. (2011). Collocated social practices surrounding photo usage in archaeology. *Computer Supported Cooperative Work, 20*, 305–340. https://doi.org/10.1007/s10606-011-9145-z

London Charter. (2006). For the use of 3-dimensional visualisation in the research and communication of cultural heritage. Retrieved August 27, 2022, from https://www.londoncharter.org/fileadmin/templates/main/docs/london_charter_1_en.pdf

London Charter. (2009). London Charter for the computer-based visualisation of cultural heritage. Retrieved August 27, 2022, from https://www.londoncharter.org/fileadmin/templates/main/docs/london_charter_2_1_en.pdf

Lopez-Menchero, V. M., & Grande, A. (2011). The principles of the Seville Charter. In K. Pavelka (Ed.), *Proceedings of XXIIIrd international CIPA symposium, Prague, Czech Republic, 12–16 September 2011*. Retrieved September 17, 2022, from https://www.cipaheritagedocumentation.org/activities/conferences/proceedings_2011/.

Lucas, G. (2001). Destruction and the rhetoric of excavation. *Norwegian Archaeological Review, 34*(1), 35–46. https://doi.org/10.1080/00293650119347

Lucas, G. (2012). *Understanding the archaeological record*. Cambridge University Press.

Lutz, C., & Collins, J. (2003). The photograph as an intersection of gazes. The example of national geographic. In L. Wells (Ed.), *The photography reader* (pp. 354–374). Routledge.

Mamassian, P., Knill, D. C., & Kersten, D. (1998). The perception of cast shadows. *Trends in Cognitive Sciences, 2*(8), 288–295. https://doi.org/10.1016/s1364-6613(98)01204-2

Miller, P., & Richards, J. (1995). The good, the bad, and the downright misleading: Archaeological adoption of computer visualization. In J. Huggett & N. Ryan (Eds.), *Computer applications and quantitative methods in archaeology 1994* (BAR international series 600) (pp. 19–22). Tempus Reparatum.

Moore, J., Rountrey, A., & Scates Kettler, H. (Eds.). (2022). *3D data creation to curation: Community standards for 3D data preservation*. Association of College & Research Libraries. Retrieved September 18, 2022, from https://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838939147_3D_OA.pdf

Morgan, C., & Wright, H. (2018). Pencils and pixels: drawing and digital media in archaeological field recording. *Journal of Field Archaeology, 43*(2), 136–151. https://doi.org/10.1080/00934690.2018.1428488

Mudge, M. (2012). Transparency for empirical data. In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.), *Paradata and transparency in virtual heritage* (pp. 252–263). Routledge.

Munsell, A. H. (1905). *A color notation*. G. H. Ellis.

Munsell, A. H. (1912). A pigment color system and notation. *The American Journal of Psychology, 23*(2), 236–244. https://doi.org/10.2307/1412843

Munsell Soil Color Charts. (1994). Macbeth Division of Kollmorgen Instruments Corporation.

Niccolucci, F., & Hermon, S. (2010). A fuzzy logic approach to reliability in heritage representation. In A. Chalmers, D. Arnold, & F. Niccolucci (Eds.), *VAST'03 proceedings of the 4th international symposium on virtual reality, archaeology, and intelligent cultural heritage* (pp. 51–60). European Association for Computer Graphics. Retrieved October 2, 2022, from https://proceedings.caaconference.org/files/2004/03_Niccolucci_Hermon_CAA_2004.pdf

Nyhan, J. (2020). The evaluation and peer review of digital scholarship in the humanities: Experiences, discussions, and histories. In J. Edmond (Ed.), *Digital technology and the practices of humanities research* (pp. 163–182). Open Book Publishers. https://doi.org/10.11647/obp.0192.07

Opgenhaffen, L., Lami, M. R., & Mickleburgh, H. (2021). Art, creativity and automation. from charters to shared 3D visualization practices. *Open. Archaeology, 7*(1), 1648–1659. https://doi.org/10.1515/opar-2020-0162

Opitz, R. S., Mogetta, M., & Terrenato, N (Eds.). (2018). *A mid-Republican house from Gabii*. University of Michigan Press. https://doi.org/10.3998/mpub.9231782

O'Shea, R. P., Blackburn, S. G., & Ono, H. (1994). Contrast as a depth cue. *Vision Research, 34*(12), 1595–1604. https://doi.org/10.1016/0042-6989(94)90116-3

Papadopoulos, C., & Earl, G. (2014). Formal three-dimensional computational analyses of archaeological spaces. In S. Polla, U. Lieberwirth, & E. Paliou (Eds.), *Spatial analysis and social spaces: Interdisciplinary approaches to the interpretation of prehistoric and historic built environments* (pp. 135–166). De Gruyter. https://doi.org/10.1515/9783110266436.135

Papadopoulos, C., & Earl, G. P. (2009). Structural and lighting models for the Minoan cemetery at Phourni, Crete. In *VAST'09, Proceedings of the 10th international conference on virtual reality, archaeology and cultural heritage* (pp. 57–64). Eurographics Association. https://doi.org/10.2312/VAST/VAST09/057-064

Papadopoulos, C., & Schreibman, S. (2019). Towards 3D scholarly editions: The Battle of Mount Street Bridge. *DHQ: Digital Humanities Quarterly, 13*(1) Retrieved September 5, 2022, from http://www.digitalhumanities.org/dhq/vol/13/1/000415/000415.html

Patrick, E. L. (1985). Is there an archaeological record? *Advances in Archaeological Method and Theory, 8*, 27–62. Retrieved October 2, 2022, from http://www.jstor.org/stable/20170186

Piccoli, C. (2016). Enhancing GIS urban data with the 3rd dimension: a procedural modelling approach. In S. Campana, R. Scopigno, G. Carpentiero, & M. Cirillo (Eds.), *CAA 2015 Keep the revolution going. Proceedings of the 43rd annual conference on computer applications and quantitative methods in archaeology* (pp. 35–44). Archaeopress.

Piggott, S. (1965). Archaeological draughtsmanship: Principles and practice. Part I: Principles and retrospect. *Antiquity, 39*(155), 165–176. https://doi.org/10.1017/S0003598X00031823

Pijpers, K. (2021). Worlding excavation practices. *Open Archaeology, 7*(1), 889–903. https://doi.org/10.1515/opar-2020-0177

Pletinckx, D. (2007). Interpretation Management. How to make sustainable visualisations of the past. In H. Gottlieb (Ed.), *EPOCH knowhow book*. Retrieved September 8, 2022, from http://repo.nodem.org/uploads/Interpretation_Managment_TII.pdf

Ramachandran, V. S. (1988). Perceiving shape from shading. *Scientific American, 259*(2), 76–83. Retrieved September 9, 2022, from https://www.jstor.org/stable/24989197

Reilly, P., Callery, S., Dawson, I., & Gant, S. (2021). Provenance illusions and elusive paradata: When archaeology and art/archaeological practice meets the phygital. *Open Archaeology, 7*(1), 454–481. https://doi.org/10.1515/opar-2020-0143

Roberts, J. C., & Ryan, N. (1997). Alternative archaeological representations within virtual worlds. In R. Boweden (Ed.), *Proceedings of the 4th UK virtual reality specialist interest group conference* (pp. 179–188). Brunel University.

Rogers, B. J., & Graham, M. (1979). Motion parallax as an independent cue for depth perception. *Perception, 8*(2), 125–134. https://doi.org/10.1068/p080125

Roussou, M., & Drettakis, G. (2003). Photorealism and non-photorealism in virtual heritage representation. In *First Eurographics workshop on graphics and cultural heritage*. Eurographics. Retrieved March 18, 2023, from https://diglib.eg.org/bitstream/handle/10.2312/VAST.VAST03.051-060/051-060.pdf?sequence=1

R & R (n.d.). *Recognition and Rewards. Room for Everyone's Talent*. Retrieved May 13, 2024, from https://recognitionrewards.nl/.

Sample, M. (2011, March 22). The poetics of metadata and the potential of paradata. *Sample Reality*. Retrieved September 18, 2022, from https://samplereality.com/2011/03/22/the-poetics-of-metadata-and-the-potential-of-paradata/

Saunders, J. A., & Backus, B. T. (2006). The accuracy and reliability of perceived depth from linear perspective as a function of image size. *Journal of Vision, 6*(9), 933–954. https://doi.org/10.1167/6.9.7

Schäfer, U. U. (2019). Uncertainty visualization and digital 3D modeling in archaeology. A brief introduction. *International Journal of Digital Art History, 3*, 87–106. https://doi.org/10.11588/dah.2018.3.32703

Schreibman, S., Mandell, L., & Olsen, S. (2011). Evaluating digital scholarship: Introduction. *Profession*, 123–135. http://www.jstor.org/stable/41714114

Schreibman, S., & Papadopoulos, C. (2019). Textuality in 3D: Three-dimensional (re) constructions as digital scholarly editions. *International Journal of Digital Humanities, 1*(2), 221–233. https://doi.org/10.1007/s42803-019-00024-6

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. MIT Press.

Sifniotis, M., Jackson, B., White, M., Mania, K., & Watten, P. (2006). Visualising uncertainty in archaeological reconstructions: a possibilistic approach. In *ACM SIGGRAPH 2006 sketches (SIGGRAPH '06).* Association for Computing Machinery. https://doi.org/10.1145/1179849.1180049

Silberman, N. A. (2008). *ICOMOS charter for the interpretation and presentation of cultural heritage sites*. ICOMOS. Retrieved September 24, 2022, from http://icip.icomos.org/downloads/ICOMOS_Interpretation_Charter_ENG_04_10_08.pdf.

Sköld, O., Börjesson, L., & Huvila, I. (2022). Interrogating paradata. In *Information Research, 27* (Special Issue). *Proceedings of the 11th international conference on conceptions of library and information science, Oslo Metropolitan University, May 29–June 1, 2022*. https://doi.org/10.47989/colis2206

Sullivan, E. A. (2020). *Constructing the sacred: Visibility and ritual landscape at the Egyptian necropolis of Saqqara*. Stanford University Press. Retrieved September 18, 2022, from https://constructingthesacred.org/.

Tarr, M. J., Kersten, D., & Bülthoff, H. H. (1998). Why the visual recognition system might encode the effects of illumination. *Vision Research, 38*(15–16), 2259–2275. https://doi.org/10.1016/S0042-6989(98)00041-8

Taylor, M. M., Lederman, S. J., & Gibson, R. H. (1973). Tactual perception of texture. In E. Carterette & M. Friedman (Eds.), *Handbook of perception III* (pp. 251–272). Academic Press.

Thomas, J. S. (1993). The politics of vision and archaeologies of landscape. In B. Bender (Ed.), *Landscape: Politics and perspectives* (pp. 19–48). Berg.

Tilley, C. (1989). Excavation as theatre. *Antiquity, 63*, 275–280. https://doi.org/10.1017/S0003598X00075992

Tilley, C. (1994). *A phenomenology of landscape. Places, paths and monuments*. Berg.

Tilley, C., & Bennett, W. (2004). *The materiality of stone. Explorations in landscape phenomenology 1*. Berg.

Tilley, C., & Bennett, W. (2008). *Body and image. Explorations in landscape phenomenology 2*. Left Coast Press.

Todd, J. T. (2004). The visual perception of 3D shape. *Trends in Cognitive Sciences, 8*(3), 115–121. https://doi.org/10.1016/j.tics.2004.01.006

Todd, J. T., Koenderink, J. J., Van Doorn, A. J., & Kappers, A. M. (1996). Effects of changing viewing conditions on the perceived structure of smoothly curved surfaces. *Journal of Experimental Psychology, 22*(3), 695–706. https://doi.org/10.1037/0096-1523.22.3.695

Topper, D. (1996). Towards an epistemology of scientific illustration. In B. S. Baigrie (Ed.), *Picturing knowledge: Historical and philosophical problems concerning the use of art in science* (pp. 215–249). University of Toronto Press.

Turner, M. J. (2012). Lies, damned lies and visualizations: Will metadata and paradata be a solution or a curse? In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.), *Paradata and transparency in virtual heritage* (pp. 210–219). Routledge.

Wade, N. J. (1999). *A natural history of vision*. Massachusetts Institute of Technology.

Westman, A. (Ed.). (1994). *Archaeological site manual* (1st ed.1980) (3rd ed.). MoLAS. Retrieved October 2, 2022, from http://www.museumoflondonarchaeology.org.uk/NR/rdonlyres/056B4AFD-AB5F-45AF-9097-5A53FFDC1F94/0/MoLASManual94.pdf

Witmore, C. L. (2004). On multiple fields. Between the material world and media: Two cases from the Peloponnesus, Greece. *Archaeological Dialogues, 11*(2), 133–164. https://doi.org/10.1017/S1380203805001479

Yarrow, T. (2003). Artefactual persons: The relational capacities of persons and things in the practice of excavation. *Norwegian Archaeological Review, 36*(1), 65–73. http://www.tandfonline.com/doi/pdf/10.1080/00293650307296

Zuk, T., Carpendale, S., & Glanzman, W. D. (2005). Visualizing temporal uncertainty in 3D virtual reconstructions. In M. Mudge, N. Ryan, & R. Scopigno (Eds.), *VAST 2005—The 6th international symposium on virtual reality, archaeology and cultural heritage* (pp. 99–106). Eurographics Association. https://doi.org/10.2312/VAST/VAST05/099-106

Costas Papadopoulos is Associate Professor in Digital Humanities and Culture Studies at Maastricht University. His work has its roots in archaeology, digital humanities, and heritage studies, exploring three-dimensional modelling and representation at the intersections of the physical and the digital. His research explores the sensorial experience and perception of heritage; engages with debates on the role of open educational resources; and explores ways to build epistemological frameworks for atypical research outputs. He is PI of PURE3D (funded by the Dutch PDI-SSH), which is developing a national research infrastructure for the publication and preservation of 3D scholarship.

Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites

Sarah A. Buchanan and Theresa Huntsman

Abstract

Archaeological fieldwork requires systematic approaches to recording and archiving a variety of data, from information about contexts and artifacts to methodologies that can vary from season to season. Drawing together retrospective analyses of data-making efforts in four student-engaged archaeological projects—the Archaeological Exploration of Sardis, American Excavations at Morgantina: Contrada Agnese Project, Poggio Civitate Archaeological Project, and Venus Pompeiana Project—this chapter focuses on the paradata work practiced by on-site data archivists. Paradata are the human processes that generate data. Field schools and the archaeological research they produce benefit from a "dusting," i.e., analytical narration of the processes, proactively led by data archivists in context and collaboration with team members. Paradata make explicit the pedagogical goals at the heart of each project while simultaneously supporting the contextual integrity of future archaeological research.

S. A. Buchanan (✉)
University of Missouri, Columbia, MO, USA
e-mail: buchanans@missouri.edu

T. Huntsman
Yale University, Yale University Office of Development, New Haven, CT, USA
e-mail: theresa.huntsman@yale.edu

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_5

1 Introduction: Archaeology as a Data Collaboration

Archaeological data archivists manage the daily recording activities of an excavation. Their work on site may encompass data entry, information creation, sense-making, legacy integration, archival research, and even ethnographic data-gathering as a participant-observer on a dig site. Such activities collectively position the data archivist as a knowledge manager—one who assembles, researches, and provides knowledge upon a collaborator's singular request. The work of answering such reference questions continually builds up the data archivist's future ability to support more complex and data-intensive knowledge production as a peer collaborator in scholarship.

Archaeology in general is increasingly recognized as a team effort. In considering two long-term excavations in Turkey and Jordan, Mickel (2016) "visualizes and measures teamwork" to connect a particular role with a level of information-sharing, revealing the underappreciated influence of such roles in the core work of "creating the archaeological record." Who the members of such teams are has also been the focus of Handley's (2015) recommendations to archaeology graduates in the years after the Great Recession of 2007–8. Handley suggests that academic programs shift focus away from the pursuit of "transferable skills" in favor of a truer kind of reflective skillset that affords one the ability to find singular meaning in the archaeological activities completed, as relevant to their identity. When read alongside the systematic review of teamwork pedagogy in higher education by Riebe et al. (2016)—which helpfully attends to the overlap of educator, student, and institutional factors—Handley's re-commitment to the ideals of socially conscious, civic responsibility outlines a bright and inclusive future for archaeological practice writ large.

Such metacognitive emphases pervade the strong ties archaeologists bolster today with military veterans (Hill, 2021) who report revived feelings of camaraderie upon completing "meticulous field work" together. The emergence internationally of veteran-focused archaeology, especially of three such projects established in 2015, is analyzed by Everill et al. (2020). They find that participation in an archaeological project resulted in both multiple improved mental health dimensions and an increased sense of value and well-being among the participants. Metacognitive aspects of archaeological practice and the resulting practice-based scholarship fundamentally shape the duration of excavations and their scholarly legacy.

Paradata, while still an emerging concept, are firmly human-centered and thus complementary to metadata, which are data-centered. This chapter defines paradata as the human processes by which a datapoint came into being. That is slightly broader than Huggett's (2014) definition of paradata as "data provenance," which can imply a single time and place of origin. Paradata allow researchers to answer how scientific data came about, just as metadata let researchers know what the core data comprise. When visualized, patterns of paradata can identify implicit problems at one of many earlier stages of data creation and offer more accurate solutions or opportunities for remediation (Choumert-Nkolo et al., 2019: 602). This chapter takes a step toward that future by introducing the “dusting,” a paradata-focused narrative from the standpoint of archaeological archiving. Four such dustings from diverse settings provide readers with experiential use cases of paradata that in combined analysis establish the concept’s fitness as a pedagogical aid—paradata are uniquely useful for teaching and learning.

1.1 Paradata Provide Pedagogical Clarity

Paradata not only arose from 3D visualization research with cultural heritage artifacts (see Papadopoulos, chapter “A Leap of Faith: Revisiting Paradata in 3D Scholarship” in this volume) but are central to the visionary guidelines recently presented by the leaders of OpenContext, an online publisher for archaeological datasets since 2006 (Kansa et al., 2020). They undergird the present writing from archaeology sites. A range of disciplines are working to move from paradata’s initial recognition as just “extra documentation” output from core data-generating activities to a more useful contribution toward accountability, traceability, tracking, and logging of permutations. Such disciplines include electronics (Gebru et al., 2021), demography (Jackson, 2017), statistics (Karr, 2010), and now archiving (Bunn & Jones, chapter “Mapping Accessions to Repositories Data: A Case Study in Paradata” in this volume; Davet et al., 2023). What are the scholarly outcomes of participants recording data in diverse ways, seen at four sites of pedagogy? This chapter’s original contribution is a people-centered reconceptualization of constraining factors at field-school sites as valuable paradata.

Before examining paradata at four diverse archaeological settings, this chapter asserts that paradata do not exist in isolation. Rather, they are one of three current types of data produced and recorded in archaeological fieldwork, each with particular usefulness for scholarship: core data, metadata, and paradata. Archaeologists place their primary focus on the raison d’être for a field season: the generation of core data in the form of physical materials recovered and their contexts, including but not limited to stratigraphic-unit characteristics, and artifact types and quantities. Secondly, archaeologists prioritize the creation of metadata, or key concepts for the organization of core data (e.g., ceramic typologies, architectural techniques, and stratigraphic categories). *Paradata*, the focus of this chapter, are data about the processes undertaken for generating both core data and metadata and are perhaps the best reflection of the underlying scholarly research priorities and motivating questions of an excavation season. Our assertion overall is that human-centered paradata recorded by archivists make engaging archaeological research possible. Paradata are a bridge between site scholars and novices.

In general, the recording of paradata tends to be given lowest priority and often remains implicit knowledge among the specific participants of a project. While that is partly due to the time-sensitive nature and limited resources/participant bandwidth of a field season, it is also due to a common reticence to show anything but “clean,” fully processed data. Without insight into the data-gathering process, there appears to be no further room or need to query said process (Kansa & Kansa, 2013).

Scribbled notes and clarifications about archaeological processes-in-place are not generally appreciated enough for the robust supplement/complement to content that they can be. In fact, publicly sharing such paradata can enliven the corresponding data. Paradata are a gateway to the core data, making ideas accessible to novice readers, entry-level archaeologists, and seasoned scholars alike. The natural and necessary limitations of data-collecting and metadata-creating efforts make paradata a prevailing support mechanism when analyzing the results of a fieldwork season. Choosing to collect one category of data can mean not collecting another category as thoroughly, and paradata can help fill that gap. Fieldwork paradata—the documentation of processes, workflows, and methodologies—can assist in the recovery of data that are desired only in hindsight, as questions and goals for archaeological analysis shift in response to evolving perspectives.

Paradata such as knowledge of team leadership and total membership (acknowledging role levels, titles, training, and experience with technologies), hourly schedules, weather conditions, and personal recording preferences such as the use of shorthand or annotations enable a data archivist to provide informed frameworks around new research questions that involve accessing and analyzing data from a given field season. The subsequent "embedding" of such knowledge among the site archaeologists, students, and broader research networks (Khazraee & Gasson, 2015, 2017) is then its own process (about which more macroscopic studies would be valuable) but one ultimately based on the objectual practices interesting to the site team.

2 Emerging Pedagogy of Field Schools

Archaeological projects that adopt a field-school structure place explicit value both on experiential learning and on pursuing longitudinal research questions central to the goal of the project. As such, each field season must focus not only on the archaeological work at hand but also on educating a new group of individuals with varying skillsets and experiences. Work necessarily moves more slowly than it would with a team of experienced archaeologists. An ongoing gap in attending to fieldwork pedagogy (Dufton et al., 2019) can leave participants to form activity "teams" under constraints of time, roles, resources, and other aspects in situ. We suggest reconsidering and reconceptualizing such constraints as paradata categories, expressly because they can provide clues or indications of what previous activity might have happened that was not documented in the way now desired. Field-school environments are most productive for considering the place of paradata in wider scholarship because they actively support interactions between data and people, especially for pedagogical purposes. (The paradata-as-pedagogy function is also examined beyond archaeology in both Dawson & Reilly, chapter "Towards Embodied Paradata. A Diffractive Art/Archaeology Approach," and Bunn & Jones, chapter "Mapping Accessions to Repositories Data: A Case Study in Paradata," in this volume.) Field schools provide a unique framework for analyzing the key role of data archivists in recording contextual information that enriches and makes accessible the findings from archaeological research.

3 Four Sites of Paradata Analyses

In the Archaeological Exploration of Sardis, American Excavations at Morgantina: Contrada Agnese Project, Poggio Civitate Archaeological Project, and Venus Pompeiana Project, the authors have led data-making activities involving diverse student teams. Over time, teams at the four sites have created data about daily trench excavations, which are captured in a variety of analog and digital ways as the following sections describe. Such a range of strategies lend the present authors insights on the wayfinding steps necessary to succeed in querying decades of legacy datasets.

Through a consideration of site-specific data-making processes, the following four subsections (dustings) surface paradata as pedagogy to support the teaching and learning of archaeological activities among project participants. Archaeological work, including data archiving, is very embodied. Replicating the able-bodied efforts of particular individual(s) who focused on such work, especially in new or challenging geographic settings, might perpetuate blind spots related to knowledge absorption and sharing. Accessibility concerns also extend to format-specific constraints and generational expectations best identified over decades. Fortunately, paradata are effective entryways for making archaeology more accessible to people of different abilities or experience levels before participating in a field season. A paradata-as-knowledge-broker approach[1] supports the inclusion of new and more contributors in a project, assisting in intellectual processes. As illustrated through the examples below, paradata can answer research questions from a trio of disciplines that intersect at field schools: archaeological curation (concerned as it is with types of artifacts and their long-term, climate-controlled preservation needs), archival science (the means of access, maintenance, and use), and archaeology (the understanding of past cultures). Each example is organized into a "dusting" or paradata-informed narrative with two parts: a description of data practices on site and an explication of the key pedagogical outcome that our archiving such paradata makes possible.

Dusting # 3a. Expert Dialogues in the Archaeological Exploration of Sardis
Data Practices: Engagement with experts and long-term team members, who are living sources of paradata, is key to successful data (re)use at both the Archaeological Exploration of Sardis, in western Turkey, and the Poggio Civitate Archaeological Project, near Siena, Italy (dusting #3c). An ongoing, joint expedition of Harvard University and Cornell University since 1958, Sardis has benefitted from consistent leadership and long-time participants (http://sardisexpedition.org/

[1] Workshop dialoguer Paul Reilly is gratefully acknowledged for this insightful phrase.

en/essays/about-sardis-expedition). Sardis was not set up as a field school, but it became (and continues to be) the training ground of dozens of archaeologists, Classicists, conservators, and more, from students at the undergraduate and graduate level to professional practitioners and university professors. All archaeological materials excavated at Sardis remain in Turkey, and all analog documentation (fieldbooks, paper drawings, etc.) returns to the permanent archive at the Harvard Art Museums in Somerville, Massachusetts. The analog resources are then digitized in the off-season and made available via a FileMaker database.

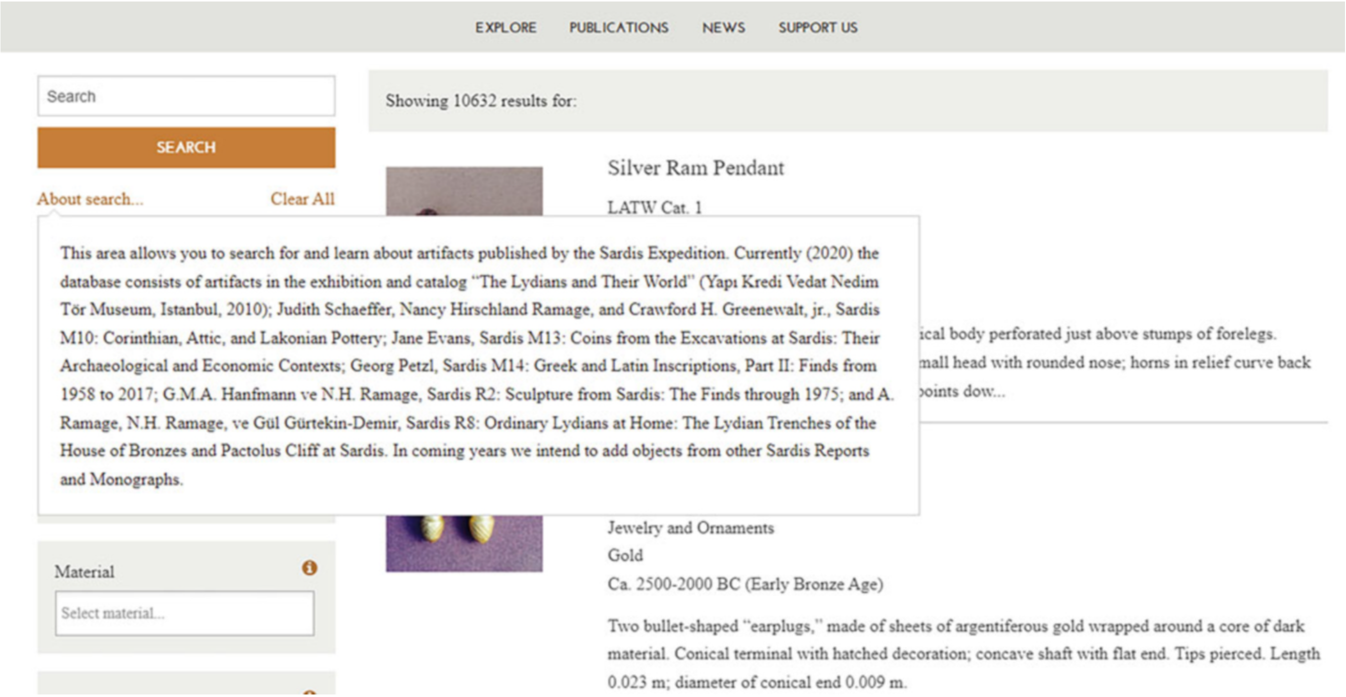
A variety of students and professionals participate in fieldwork at Sardis, and the data they generate requires different but interrelated parameters for recording and publication. The structure of the current FileMaker database reflects the original analog data-recording system established in 1958, as born-digital data must be integrated with a substantial amount of legacy data. Work on site today requires a combination of analog and digital recording during the field season itself and then the digitization/digital preservation of analog content throughout the rest of the year. Students are involved in all phases. All team members and researchers receive basic database training, but there are only a set number of individuals who have permission to make updates to records (e.g., core staff like the registrar, archivists, and conservators).

Outcomes from Paradata: As Publications Data Manager from 2014 to 2018, Huntsman contributed to the launch of and content development for Sardis's first public-facing website (http://sardisexpedition.org). The site was part of a push to make data more accessible to the general public and to expand the reach of traditional print publications. That process revealed an overwhelming need to consult Sardis's living sources of paradata. Like many twentieth-century archaeological projects, the explicit, systematic recording of paradata was not part of the data-generating process. Instead, Sardis has been able to rely on staff members deeply familiar with the history of the site to fill knowledge gaps.

Publication drives digitization priorities at Sardis, given the massive amount of information recorded, and publications often involve materials excavated anywhere from 1 to 60 years ago. The preparation of data-heavy manuscripts in which Huntsman was involved (e.g., Evans, 2018; Petzl, 2019; Ramage et al., 2021) revealed confusing and incomplete artifact records that could only be resolved with the insights of team members. Through that process of resolution, Sardis has generated and continues to create a considerable amount of paradata—often, narrative explanations as to which resources were consulted to resolve a data discrepancy, how past archaeological methods or circumstances affected it, who was involved, and when the solution was implemented.

While such paradata are important for internal purposes and critical to the accurate publication of materials, they do not necessarily belong in the published record. What does that mean for the presentation of more and more data on the Sardis website? Under the search function is an explicit acknowledgment that the artifact data presented is incomplete (Fig. 1).

The artifacts description employs transparency as a strategy to engage and involve experts in the enhancement (even co-creation) of accurate metadata that is



Showing 10632 results for:

Silver Ram Pendant

LATW Cat. 1

…ical body perforated just above stumps of forelegs. …mall head with rounded nose; horns in relief curve back …points dow...

This area allows you to search for and learn about artifacts published by the Sardis Expedition. Currently (2020) the database consists of artifacts in the exhibition and catalog "The Lydians and Their World" (Yapı Kredi Vedat Nedim Tör Museum, Istanbul, 2010); Judith Schaeffer, Nancy Hirschland Ramage, and Crawford H. Greenewalt, jr., Sardis M10: Corinthian, Attic, and Lakonian Pottery; Jane Evans, Sardis M13: Coins from the Excavations at Sardis: Their Archaeological and Economic Contexts; Georg Petzl, Sardis M14: Greek and Latin Inscriptions, Part II: Finds from 1958 to 2017; G.M.A. Hanfmann ve N.H. Ramage, Sardis R2: Sculpture from Sardis: The Finds through 1975; and A. Ramage, N.H. Ramage, ve Gül Gürtekin-Demir, Sardis R8: Ordinary Lydians at Home: The Lydian Trenches of the House of Bronzes and Pactolus Cliff at Sardis. In coming years we intend to add objects from other Sardis Reports and Monographs.

Jewelry and Ornaments

Gold

Ca. 2500-2000 BC (Early Bronze Age)

Two bullet-shaped "earplugs," made of sheets of argentiferous gold wrapped around a core of dark material. Conical terminal with hatched decoration; concave shaft with flat end. Tips pierced. Length 0.023 m; diameter of conical end 0.009 m.

Fig. 1 Screenshot of "About search" information on "Explore" subpage of Sardis website (sardisexpedition.org). ©Archaeological Exploration of Sardis/President and Fellows of Harvard College

more useful to scholars than the version without their input. Additionally, related approaches detailed above are employed both during and after the field season to further encourage data (re)use: e.g., during the season, the data archivist may translate or otherwise boost the legibility of data so members of the team—with varying levels of familiarity with the site—find it accessible/intelligible, and provide narratives to outside researchers when necessary. After each season, Huntsman and now other staff codify those narratives and present them alongside core data, regularly obtaining outside feedback on the same narratives to ensure intelligibility.

Sardis's rich data archive is accompanied by an equally rich human archive that enhances, teases out, and sheds light on old and new information. Team members from undergraduates to seasoned professionals have the opportunity to interact with living legacies on a daily basis and now are encouraged to add paradata to records when a complex problem could lead to the loss of information in the archaeological process. Data archivists stateside work in collaboration with the field team, guiding students and researchers in the most effective ways to use the database, and how to know when they have exhausted all sources of information to resolve a problem. All team members rely on the database, and those directly involved in maintaining the archive must be able to provide the guidance and clarity those individuals need for their work. In this way, the living archive becomes accessible and intelligible to all, leading to more comprehensive data and more accurate research conclusions.

Dusting # 3b. Data Journeys Across Teams and Time in the American Excavations at Morgantina: Contrada Agnese Project

Data Practices: Morgantina is an ancient Greek site in central Sicily, Italy, which has been formally excavated since 1955, then under the direction of two Princeton University professors and with contributions by Swedish archaeologists, architects, photographers, and royalty. The site has been incredibly prominent in the training and careers of American archaeologists (including the first director at Poggio Civitate; see dusting #3c), who together comprise an "archaeological family tree with some of the most respected names in twentieth century classical studies in the United States" (Edlund-Berry & Kyllingstad, 2018: 2). Contrada Agnese, the western area of the ancient city, was partially excavated from 1971, and the Contrada Agnese Project (CAP) commenced in 2013 (Schirmer et al., 2021; Walthall, 2021). The CAP team structure is both established and flexible—two aspects which lend support to the concept and practice of "data journeys" (Bates et al., 2015a, 2015b). As of 2015, when Buchanan participated in its third field season, CAP comprises the Data, Dig, Museum (Small Finds, Ceramics, Conservation, Environmental), Geospatial, and Architecture teams whose members work together to identify the types of information gathered in their work and the optimal ways of accessing information to carry out daily activities (Smalling et al., 2017). Small Finds team members, for example, capture digital photographs of excavated objects before and after conservation treatment, and their images are then linked to new object records and trench contexts in the database. The five teams have evolved over time, reflecting wider adoptions across several American excavations of digital data-recording practices after 2010. For example, archaeologists Wallrodt et al. (2015)

describe their experience of having introduced Apple iPads for recording excavation information at Pompeii during 2010–13, finding that same-day data processing enabled more sophisticated decision-making in the field.

The CAP database started with the project in 2013 as a way of recording finds and has been expanded to include records for seven types of artifact-related data: Inventoried Finds ("museum" or cataloging work), Contexts (fieldwork), Media (photographs), Science (soil and paleobotanical results), Pottery (ceramics), Storage (locations), and Conservation (treatments). Based on the individual or team, the Data Supervisor prioritizes the layout view of the database that is most useful for the team member's particular use, aware of the problem of "data overload . . . present[ing] the right data to the right audience at the right time" (Data Supervisor, 22 June 2015 fieldnotes). The Data Team ensures that all members of the excavation are able to use the data efficiently: "That's kind of where [Data Team members] come into play, in terms of the data curation: making sure that it's user-friendly, and that it makes sense, and that it's stored where it needs to be stored, and that it's easily accessible to everyone when they need it—either during the season or during the off-season" (Data Supervisor, 13 February 2015 pre-season interview). The Data Supervisor engaged in a process throughout the season of responding to trench supervisors' feedback and reactions from their use of the database and made enhancements to the database layout accordingly. One outcome of the database's iterative development was the idea to present a database workshop during the first week of the 2016 season so that all team members would see the equipment and receive instruction for particular recording activities. Additionally, a CAP Field and Museum Handbook was drafted in the months before the 2015 season.

Outcomes from Paradata: Departing from early interest in a time-use study on taking notes in a notebook or on an iPad, the Data Team ultimately refined the focus of inquiry from quantification of time and work rates to a more qualitative focus on excavators' uses of both analog and digital recording tools. To support the inquiry, five trench supervisors provided individual perspectives regarding their particular uses for paper and digital technologies. Findings showed that the paper notebook is especially useful in capturing developing narratives, sketches (of things and ideas), and chronologies and that the iPad is especially useful in capturing and sharing the final interpretation of record. Some of the enhancements suggested for the database were a place for a checklist/running tally space (for Finds and Contexts), a better linkage with iPhoto, and a functionality like Pages for journaling and sketching. The researcher asked whether using a notebook and an iPad is an either/or situation (i.e., is one nonessential?), and a supervisor replied that "it is getting close" but not there yet. Technologies range from the humble sketch in colored pencil to the high-resolution photosynth/3D model of a trench feature on a sunny day, each supporting the mission of the project and promoting archaeological research. Continued reflection by directors on the relative contributions of each technology to site knowledge will better equip all collaborators to manage archaeological data, and encourage its use.

Overall, data management on site as practiced in CAP's field-school pedagogy can extend research goals further: the information scientist working to curate and provide access to datasets, the archaeologist working to enter enough and sufficiently granular data to detect assemblage patterns (Huvila, 2014), and the community member working to research the provenance of an artifact to interpret it in a museum exhibit. Each perspective brings something to the archaeological endeavor and reveals a multiplicity of designs on the data, during and after their initial gathering. The format-specific field documentation practices of interest here remain an active area of research and development (Huvila, 2015; Morgan et al., 2021).

Dusting # 3c. Accountability Embedded in the Poggio Civitate Archaeological Project

Data Practices: Like Sardis (dusting #3a), the Poggio Civitate Archaeological Project is a long-running excavation that has also benefitted from consistent leadership. It also has an astounding amount of legacy data, most of which have been digitized since the implementation of a bespoke SQL database in 2001 (though digital efforts with FileMaker began in 1997). Poggio Civitate began with much the same team structure as Sardis: graduate students and early-career scholars directing teams of local workers, recording everything in fieldbooks, cataloging objects on typed notecards, paper architectural plans, and black-and-white photography. In the late 1970s, Poggio Civitate shifted into the field-school model (which continues today), with groups of undergraduate students earning college credits for participating in the summer excavation season. Students engage in all aspects of work while also conducting research, writing papers, visiting local museums, and listening to lectures from staff members.

Huntsman joined the team as a student in the 2001 season, when all staff and students took rotations doing data entry for all paper catalog cards into the new database, as well as transcribing fieldbooks. The digitization process continued through 2005 and included changes to the database along the way as the team encountered shifts in data-recording practice over the decades of finds. Instead of selected published materials like the Sardis website, the goal of the Poggio Civitate website was to present a full, publicly accessible and annually updated version of the entire archive, with the artifact catalog and accompanying contexts via the fieldbooks at its core.[2] This process allowed all field-school students to be directly involved in the creation of a digital archive, helped them understand the complicated nature of legacy data, and inspired conversations as to how to represent this data clearly to future researchers and the public.

As a member of the team throughout this first push for digitization and beyond, acting registrar and cataloger Huntsman tracked and documented a considerable amount of paradata in her master's thesis (Huntsman, 2005), including the decision-making process among project directors, database administrators, and other field-

[2] The original version of the database first made available online is no longer accessible, and a partnership with the linked open data initiative OpenContext is ongoing; see http://poggiocivitate.com/excavation-database.

school staff members for addressing missing data and types of data not recorded in earlier years of excavation. Such work also relied upon insights from interviews with early members of the project.

Outcomes from Paradata: In the interest of presenting the complete archive online, paradata explaining missing contextual information for an artifact, for example, were added consistently during the catalog digitization process. When there are no precise coordinates for the findspot of an artifact from a particular trench, or if findspot information cannot be rectified with the current grid, a note is added with all available information that can be gleaned from the fieldbook. That helps any viewer know why an artifact cannot be plotted on the site plan, and it also helps data archivists to know which resources were consulted.

Poggio Civitate began to work with web-based research data management and publishing service OpenContext in 2011 on creating a sustainable digital archive that could handle the problems caused by ever-changing software requirements for the earlier iteration. Analog *and* digital legacy data and paradata were key to the conversation, as were technological, archaeological, and archival perspectives. This “second round” of dealing with paradata in the creation of a digital archive reaffirmed the importance of documenting decision-making processes not only in the database itself but also in other forms (Huntsman & Kansa, 2016). Now at Poggio Civitate, students are familiarized with not only the database but how it was created, alongside the paper archive of catalog cards and fieldbooks that remains in the lab today.

Writing a narrative of the (data archivist’s) legacy research process is an exceedingly productive activity, as it builds both transparency and accountability into the data (re)use activity and promulgates an awareness of archival labor where and to whom (in this case, the researcher) it is most valuable. This is particularly important for the Poggio Civitate data on OpenContext, a linked open data platform that encourages making new connections across sites, regions, and collections. The recording of paradata is now an integral part of pedagogy and practice at Poggio Civitate and will have an impact on research not only about the site itself but on the work of those consulting data from across the ancient world on OpenContext.

Dusting # 3d. Collective Discovery in the Venus Pompeiana Project

Data Practices: Located at the southwest corner of the ancient city of Pompeii, Italy, the sanctuary site dedicated to Venus (the Roman goddess of love) has been excavated intermittently since 1898, with recent periods of fieldwork in the 1990s and mid-2000s. The international collaboration known as the Venus Pompeiana Project (VPP) commenced in 2017 with a goal of clarifying the chronology, development, and nature of occupation of the site (Battiloro & Mogetta, 2017). Extensive data are generated through such varied methods as digital photogrammetry, artifact illustrations, special-find photographs, feature sketches, bulk-find tabulations, and digital fieldnotes on the site. Established for the 2017 season by staff leaders and maintained ever since, a FileMaker database serves to document six components (stratigraphic units, photos, finds, masonry, and revetments). From its active upkeep to its uses year-round for academic study, the VPP database exemplifies intellectual discovery made collectively rather than individually.

The 2017 and 2018 seasons saw student participants work with staff leaders to successfully enter data while on site. The data archivist's addition to the team in 2019 allowed for a more embedded and dedicated approach to data entry. Following four excavation days of observation and computer setup that year, the project implemented a "data station" whereby one student participant would rotate daily into the data-archiving activity, working closely with the data archivist. The archivist and student together discussed and agreed upon priorities and processes to be completed for the day, in conversation with trench and pottery supervisors and project directors. Activities completed by the two data-station team members varied depending on the progress of unit documentation drafting and collaborative review, the intensity of excavation activities in the trenches, and sometimes logistical hurdles related to the availability of Internet signal and/or particular field notebooks with essential information for the database.

As part of a funded effort to present the project's "digital dig" data and narrative accompanying its forthcoming publication phase, project leaders established a data curation collaboration with OpenContext so that component information can be shared and preserved, and designed a project website to support storytelling and an artifact-embedded virtual tour of the site (https://venuspompeiana.mused.org/; https://www.archaeological.org/interactive-dig/pompeii-italy/). The Mused website greatly expanded during 2020 when the COVID-19 pandemic delayed fieldwork for two seasons; during that time, the team collaboratively posted six narrative topic pages and over two dozen artifact descriptions with photographs.

Outcomes from Paradata: In preparation for the 2022 field season, expansion of the data archivist's role from one to two positions was generously accommodated. Such recognition by the project directors, staff leaders, and all student participants of the value of data archiving concurrent with site work strengthened the team's collective ability to stay abreast of fast excavation progress in all open trenches and support specialized analysis of artifact subcategories. As one measure of that progress, Fig. 2 visualizes the amount of Find Numbers assigned during respective seasons of the project. The project's nearly two hundred small finds each have catalog records and one or more photographs linked in the database. Relatedly, the VPP employs a recovery strategy in the field of total dry screening by stratigraphic unit. The total volume of deposits excavated each year (not including any backfill cleared), per the VPP topographer's geodatabase, is approximately 33 cubic meters (m^3) in 2017, 66 m^3 in 2018, 20 m^3 in 2019, and 33 m^3 in 2022.

The Venus Pompeiana Project is distinguishable by the immediacy and integration of archiving with the fieldwork. Its model of knowledge production is dynamic: a genuine give-and-take with a great degree of listening to the multiple perspectives gathered together on site. Like CAP (dusting #3b), VPP has a small and well-defined scope and was born a hybrid project (with paper and digital operations), allowing it to be more agile with its archive. Such agility means that a new data field can be added and researched for use with reasonable flexibility. The six components recorded in the database include initials of those who entered the data for each

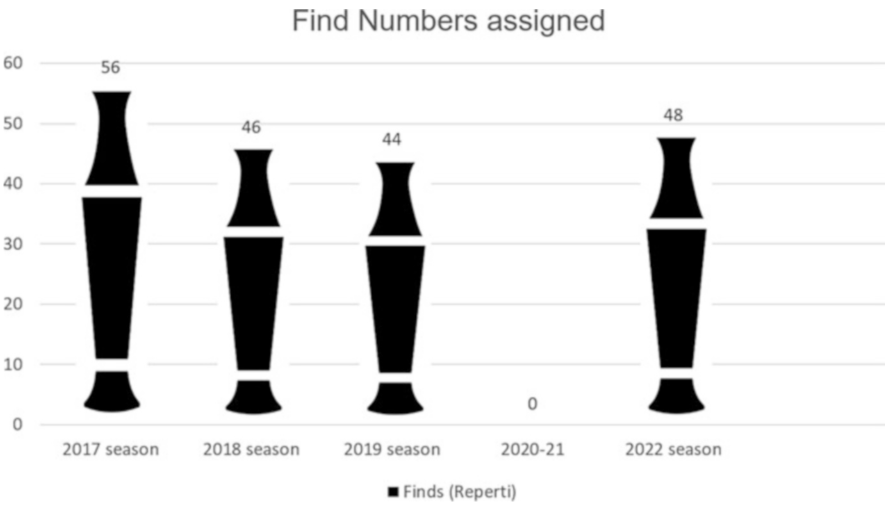


Fig. 2 Number of finds assigned during field seasons

record, which itself exists within a defined scope such as a given trench or data category. Paradata continue to be key to the project's successful data archive, and its mindful permeation into the work of two recent field seasons is something that was not the case at sites like Poggio Civitate or Sardis (dustings #3c and 3a) until the present time, as contributors realize that their digital data are growing. Paradata aspects should continue productively shaping the archiving and recording of archaeological data.

4 Presence and Proximity: An Integrative Discussion on Using Paradata Narratives (Dustings) to Support Instruction on Dig Sites

While a consideration of the four sites above contributes a new layer of paradata-based knowledge about the archaeological progress made across a given fieldwork season, the paradata narratives also have practical usefulness. The additive writing of and about processes behind the data studied by field archaeologists illuminates habits and patterns of work that take root during a season and can become routinized to the point of discomfort at alternate ways or alternate performers of that work. Bringing those patterns to the surface during post-season team debriefs (and even long after, which is far rarer) is an important responsibility and contribution of a data archivist.

Since the data archivist very often is on the move collecting data from trenches, labs, visitor audiences, and/or site directors in the course of both a single day and a set of dig days, they have unique knowledge of how individual team-member contributions make up the collective progress, and they could help qualitatively evaluate which efforts were successful, efficient, or effective learning experiences. Their co-presence and co-commitment to more than one expert positions them to lead the important work of integrating data from multiple contributors, sharing that knowledge with the team, and ultimately presenting a coherent narrative about the fieldwork findings.

Even so, paradata as an exploratory concept (at VPP in 2022 or the other sites discussed here) does not call for an extra, added-on task of collecting more new data. Indeed, paradata are not necessarily recorded as such in legacy materials. Paradata from work prior to VPP's start in 2017 instead exist in the form of two individuals who were present for earlier campaigns and involved in the current project (Ivan Varriale and Marcello Mogetta). Because the Temple of Venus at Pompeii is a site that has been and could be excavated multiple times, it is prudent for archaeologists to ensure that paradata enter the archive in some form. At long-running projects like Sardis and Poggio Civitate, that pressure is somewhat lessened due to their status as continuous digs, which benefit from deep knowledge transfer happening among long-time leaders and have no set project end dates. Yet VPP, like CAP, is a discrete operation that has some born-digital components coexisting in a longer, analog, inherited history of site excavation. While mindful of that recency (being an impetus of sorts), VPP and CAP still incorporate paradata into their excavation operations and archives. Paradata are found to be ultimately well suited for both the digital aspects of modern archaeological excavation and their effective presentation to today's audiences seeking out the historic legacies of known, storied sites.

The natural continuation of a paradata-integrative operation (as opposed to the codification of paradata well after the fact) such as at VPP can result in substantive benefits to the project. The mere presence of data archivists on site can ensure that if and when some data are found to be incomplete, the data question can be asked in the moment, and individuals can be kept informed as to the answer. Furthermore, the proximity of the archivist alleviates problems and mistakes that otherwise would be difficult to resolve when discovered later. Usually such issues stem from a minor mistake or mishearing. Rather than compounding, they can be prevented.

To be clear, such initiative exists beyond any one person or station, and it is the presence of someone who is attuned to and aware of paradata about the scope of the operation and what data work has happened in the past. Such individuals can advise decision-making in a productive way. Students witnessing that dialogue thus become encouraged to talk issues out with their teammates and problem-solve together, modeling exemplary collaborative behavior while practically keeping interested parties in the loop. Project participants create partnerships on a respectful playing field. Furthermore, such an environment reinforces the value and necessity of diverse experiences and training backgrounds which generate different questions. In turn, these insights make the processes in use better for the new purposes. Instead of locking new participants into tracks or static workflow patterns, the paradata actors help set up students as active contributors in making the work better. Across the cohorts of each project season, there are naturally some who gravitate confidently toward an evolution of work mindset and others who prefer to receive instruction and maintain a status quo—perhaps all such contributions reflect the diverse education that shape one's very personal reasons for wanting to participate. Again, all forms and means of contributing to data-archiving efforts such as paradata-recording facilitate its permeation well beyond the archivist so that the operation is sustainable.

5 Conclusion

A paradata-aware approach to data archiving on a dig team illuminates patterns, helps separate individual habits from the role of archiving, supports attribution, fosters dialogue and teamwide problem-solving, and generates visualizations of the archiving process and products. Archaeological field seasons bring a group of people together for an intense experience where everyone develops some shared understanding of and terms for the work at hand as a matter of teamwork. Many projects operate under the general assumption that "everybody knows that X term in the data means Y modification for the duration of the present conditions." A successful season will generate a critical mass of people "in-the-know" in ways they were not before, but the archival perspective reveals that is not always lasting, especially over the course of many years, or even between two consecutive seasons of work. The experience-sharing can feel exclusionary—to someone who did not know some background information or who missed witnessing a particular situation—and can deter a laudable attempt to solve problems that arise. (Temporary) knowledge production in such settings occurs subtly and quietly, such that only at times of need would such participants be able to acknowledge being "out of the groove" and in a position, perhaps, to seek the information. At that point, however, the information or datapoint might be lost; whether it is big or small in size, its absence may eventually affect the content, quality, or utility of data recorded.

Recording such data in ways that are intelligible to future researchers is critical. Archaeology is a destructive science by its very nature, so supporting replicability and accessibility are essential aims of data archiving. Paradata shed light on issues of temporality and ephemerality that have always been present on archaeological projects, by illuminating the sense of immediacy that such projects generate—that "it will always be this way." Even so, paradata are not exclusive to archivists, integrated as one may be on an excavation. Instead, they touch upon all-too-human episodes, incidents, and characteristics that go unspoken for one reason or another. The capable embodiment of a role should not forever tie those particular behaviors to that (archiving) role but, rather, attentiveness to such paradata issues should encourage the tasks-of-the-role's permeation to the other participants, since fieldwork is a collaborative endeavor. A thoughtful smattering of details and reasons why data appear the way they do enlivens an accompanying narrative and connects us all as human beings.

References

Bates, J., Goodale, P., & Lin, Y. W. (2015a). Data journeys as an approach for exploring the socio-cultural shaping of (big) data: The case of climate science in the United Kingdom. In *iConference 2015 preliminary results papers*. Retrieved September 3, 2022, from http://hdl.handle.net/2142/73429.

Bates, J., Goodale, P., & Lin, Y. W. (2015b). Mapping data journeys: Design for an interactive web site. In *iConference 2015 poster descriptions*. Retrieved September 3, 2022, from http://hdl.handle.net/2142/73757.

Battiloro, I., & Mogetta, M. (2017). Tempio di Venere. *Fasti Online*. Retrieved September 3, 2022, from https://www.fastionline.org/record_view.php?fst_cd=AIAC_811.

Choumert-Nkolo, J., Cust, H., & Taylor, C. (2019). Using paradata to collect better survey data: Evidence from a household survey in Tanzania. *Review of Development Economics, 23*(2), 598–618. https://doi.org/10.1111/rode.12583.

Davet, J., Hamidzadeh, B., & Franks, P. (2023). Archivist in the machine: Paradata for AI-based automation in the archives. *Archival Science*. https://doi.org/10.1007/s10502-023-09408-8.

Dufton, J. A., Gosner, L. R., Knodell, A. R., & Steidl, C. (2019). Archaeology underfoot: On-campus approaches to education, outreach, and historical archaeology at Brown University. *Journal of Field Archaeology, 44*(5), 304–318. https://doi.org/10.1080/00934690.2019.1605123.

Edlund-Berry, I., & Kyllingstad, R. (2018). Morgantina revisited: An architect's recollections. *CLARA: Classical Art and Archaeology, 3*. Retrieved September 3, 2022, from https://doi.org/10.5617/clara.v3i0.6064.

Evans, J. D. (2018). *Coins from the excavations at Sardis: Their archaeological and economic contexts, coins from the 1973 to 2013 excavations*. Sardis Monograph 13. Retrieved September 3, 2022, from https://sardisexpedition.org/en/publications/m13.

Everill, P., Bennett, R., & Burnell, K. (2020). Dig in: An evaluation of the role of archaeological fieldwork for the improved wellbeing of military veterans. *Antiquity, 94*(373), 212–227. https://doi.org/10.15184/aqy.2019.85.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723.

Handley, F. (2015). Developing archaeological pedagogies in higher education: Addressing changes in the community of practice of archaeology. *The Historic Environment: Policy & Practice, 6*(2), 156–166. https://doi.org/10.1179/1756750515z.00000000080.

Hill, M. (2021). *Veterans at revolutionary battlefield dig find camaraderie*. Associated Press. Retrieved January 3, 2023, from https://apnews.com/article/lifestyle-health-veterans-post-traumatic-stress-disorder-science%2D%2D7ae7c1b4a5eecdc843f01d2c34c89f69.

Huggett, J. (2014). Promise and paradox: Accessing open data in archaeology. In C. Mills, M. Pidd, & E. Ward (Eds.), *Proceedings of the digital humanities congress 2012: Studies in the digital humanities*. Digital Humanities Institute. Retrieved January 3, 2023, from https://www.dhi.ac.uk/books/dhc2012/promise-and-paradox/.

Huntsman, T. (2005). *Excavating the catalog: A study of cataloguing practice at Poggio Civitate (Murlo)*. MA thesis, Washington University in St. Louis.

Huntsman, T., & Kansa, E. C. (2016). *Poggio digitate: The history and future of data recording and presentation at an Etruscan site*. Presented at the 117th annual meeting of the Archaeological Institute of America, January 6–9, 2016, San Francisco, California.

Huvila, I. (2015). *Chatting #fieldnotes: Rethinking notetaking workflows in field archaeology*. Poster presented at computer applications and quantitative methods in archaeology (CAA) annual conference, Siena, Italy. Retrieved September 3, 2022, from http://istohuvila.se/node/440.

Huvila, I. (2014). Archaeologists and their information sources. In I. Huvila (Ed.), *Perspectives to archaeological information in the digital society* (pp. 25–54). Institutionen för ABM. Retrieved September 3, 2022, from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-240334.

Jackson, H. (2017). What are paradata?—An example. *Research Matters: CENSUS Blog*. Retrieved January 3, 2023, from https://www.census.gov/newsroom/blogs/research-matters/2017/04/paradata.html.

Kansa, E. C., & Kansa, S. W. (2013). We all know that a 14 is a sheep: Data publication and professionalism in archaeological communication. *Journal of Eastern Mediterranean Archaeology and Heritage Studies, 1*(1), 88–97. https://doi.org/10.5325/jeasmedarcherstu.1.1.0088.

Kansa, S. W., Atici, L., Kansa, E. C., & Meadow, R. H. (2020). Archaeological analysis in the information age: Guidelines for maximizing the reach, comprehensiveness, and longevity of data. *Advances in Archaeological Practice, 8*(1), 40–52. https://doi.org/10.1017/aap.2019.36.

Karr, A. F. (2010). *Metadata and paradata: Information collection and potential initiatives. Expert panel report*. National Institute of Statistical Sciences. Retrieved January 3, 2023, from https://www.niss.org/research/metadata-and-paradata-information-collection-and-potential-initiatives.

Khazraee, E., & Gasson, S. (2015). Epistemic objects and embeddedness: Knowledge construction and narratives in research networks of practice. *The Information Society, 31*(2), 139–159. https://doi.org/10.1080/01972243.2015.998104.

Khazraee, E., & Gasson, S. (2017). Epistemic objects and embedded paradigms. *Academy of Management Proceedings, 2014*(1). https://doi.org/10.5465/ambpp.2014.15832abstract.

Mickel, A. (2016). Edges of teamwork in archaeology: Network approaches to excavation histories. In *Society for American Archaeology 81st annual meeting, Orlando, FL* (tDAR id: 404480). tDAR. https://doi.org/10.6067/XCV8KS6T9R.

Morgan, C., Petrie, H., Wright, H., & Taylor, J. S. (2021). Drawing and knowledge construction in archaeology: The Aide Mémoire Project. *Journal of Field Archaeology, 46*(8), 614–628. https://doi.org/10.1080/00934690.2021.1985304.

Petzl, G. (2019). *Sardis: Greek and Latin Inscriptions Part II, Finds from 1958 to 2017*. Sardis Monograph 14. Retrieved September 3, 2022, from https://sardisexpedition.org/en/publications/m14.

Ramage, A, Ramage, N. H., & Gül Gürtekin-Demir, R. (2021). *Ordinary Lydians at home: The Lydian trenches of the House of Bronzes and Pactolus Cliff at Sardis*. Sardis reports 7. Retrieved September 3, 2022, from https://sardisexpedition.org/en/publications/r8.

Riebe, L., Girardi, A., & Whitsed, C. (2016). A systematic literature review of teamwork pedagogy in higher education. *Small Group Research, 47*(6), 619–664. https://doi.org/10.1177/1046496416665221.

Schirmer, C., Walthall, D. A., Tharler, A., Wueste, E., Crowther, B., Souza, R., Benton, J., & Millar, J. (2021). Preliminary report on the 2018 field season of the American Excavations at Morgantina: Contrada Agnese Project (CAP). *Fasti Online*, 500. Retrieved September 3, 2022, from http://www.fastionline.org/docs/FOLDER-it-2021-500.pdf.

Smalling, A., Lieberman, L. A., Truetzel, A. E., & Alex Walthall, D. (2017). Facilitating collaboration between archaeologists and conservators at Morgantina (Sicily). In N. Owczarek, M. Gleeson, & L. A. Grant (Eds.), *Engaging conservation: Collaboration across disciplines* (pp. 90–97). Archetype Publications.

Wallrodt, J., Dicus, K., Lieberman, L., & Tucker, G. (2015). Beyond tablet computers as a tool for data collection: Three seasons of processing and curating digital data in a paperless world. In A. Traviglia (Ed.), *Across space and time: Papers from the 41st annual conference of computer applications and quantitative methods in archaeology (CAA), Perth, 25–28 March 2013* (pp. 97–103). Amsterdam University Press. https://doi.org/10.5117/9789089647153.

Walthall, D. A. (2021). The Hellenistic house in motion: Reflections on the CAP excavations at Morgantina (2014–2019). In C. Prescott, A. Karivieri, P. Campbell, K. Göransson, & S. Tusa (Eds.), *Trinacria, 'An island outside of time': International archaeology in Sicily* (pp. 55–70). Oxbow Books. https://doi.org/10.2307/j.ctv24q4z4h.12.

Sarah A. Buchanan is an associate professor at the University of Missouri serving as lead faculty in Archival Studies. Her 2016 dissertation investigated archaeological curation in the field and for storytelling purposes, and subsequent work has focused on provenance research methods and the preservation of audiovisual collections. Informed by prior experience as a museum archivist and librarian, she teaches honors undergraduate and graduate courses in information science and is active in the Society of American Archivists as a student chapter faculty advisor.

Theresa Huntsman received her PhD from Washington University in St. Louis in 2014 with her dissertation, "Eternal Personae: Chiusine Cinerary Urns and the Construction of Etruscan Identity." She has robust experience in researching and assisting in the preservation, presentation, and publication of archaeological data, including for the Poggio Civitate Archaeological Project (2001–13) and the Archaeological Exploration of Sardis (2014–18). She now serves as Associate Director of Special Stewardship at Yale University and collaborates with archaeological projects and collecting institutions through data management and editorial work.

Towards Embodied Paradata. A Diffractive Art/Archaeology Approach

Ian Dawson and Paul Reilly

Abstract

For archaeologists, artists, and cultural heritage workers, paradata are generally viewed as explicitly *selected* and *documented* attributes, or defined sets of circumstances, authoritatively considered to have a material outcome on the provenance, collection, and manipulation of both recorded data and metadata and their subsequent interpretation or analysis of artefacts and other (contextual) remains. Being chosen, their own provenance is questionable: why were the selected data, metadata, and paradata more relevant than other options? We (re)consider embodied practice as a form of paradata-making normally airbrushed out of the hegemonic accounts of how works of art and archaeological excavations are presented and analysed. Decisions to not include the embodied worker, their apparatus, and their practices of making, or uncovering, haunt images purporting to be historical accounts in the art and archaeology literature by their absence. Adopting a diffractive art/archaeology approach, and subversively applying several well-known cultural heritage recording and presentation techniques, recursively and unconventionally, we throw light on embodied paradata and (re)present them as potentially very valuable pedagogical boundary objects. We also dislocate paradata away from a purely epistemological dimension into an entangled onto-epistemological nexus.

I. Dawson
Winchester School of Art, University of Southampton, Winchester, UK
e-mail: i.dawson@soton.ac.uk

P. Reilly (✉)
Faculty of Humanities, University of Southampton, Southampton, UK
e-mail: p.reilly@soton.ac.uk

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_6

1 Introduction and Background

Artefacts and artworks are widely supposed to stand up and speak for themselves. Opinions appear to be ancillary. We believe this is misleading. Artefacts and assemblages are deliberately articulated in ways to present specific points of view. Transdisciplinarity or cross-boundary knowledge sharing—that is reading through one another's disciplinary filters—offers the promise of revealing informative diffraction patterns in which these invisible decision-making processes, choices, and opinions become visible and where productive knowledge accidents might appear. Promising, but no one said it would be easy. Transdisciplinary knowledge sharing, management, and communication are without doubt challenging. The accounts we offer for the decisions we make, in other words the paradata we feature, in practice-based disciplines such as field archaeology and fine art, where much knowledge is tacit and unspoken, are not necessarily readily apparent, especially for newcomers or students to the field. Even where we have developed explicit disciplinary conceptual frameworks, miscommunication across disciplinary or domain borders is a real possibility. For instance, the words *provenance*, *provenience*, *paradata*, *context*, and *assemblage* are examples of shared terminology for a sophisticated group of related concepts that, allegedly, account for the circumstances of discovery, interpretation, and subsequent life history of an artefact or assemblage and have subtle but consequential differences in meaning in the disciplines of art, archaeology, and cultural heritage more generally (Huvila & Sköld, 2021, chapter 'A Leap of Faith: Revisiting Paradata in 3D Scholarship' in this volume; Reilly et al., 2021; Sköld et al., 2022). In this chapter, we focus on embodied paradata.

Our approach to embodied decision-making is from an *Art /Archaeology* perspective, which is not a simple combination of art in archaeology or archaeology in art. We adopt a far more disruptive and unsettling emerging discipline which constitutes a kind of diffraction zone where *embodied,* and *material*, archaeological, and artist perspectives and practices interlace one another (Bailey, 2017, 2018). The resulting interference patterns that emerge in this transdisciplinary diffraction zone show important, perhaps subversive, but revealing aspects of difference.

Both authors were trained in distinct practice-based disciplines. Our knowledge continues to build through embodied experiences of interacting with materials in trenches and studios, where embodied practices intersect and mingle in the realms of (im)materiality, temporality, movement, gesture, and mark-making. Dawson is a sculptor. Reilly is an archaeologist. We are also collaborators entangled with digital technology, particularly technologies of imaging and 3D printing. We explore diffractively the affordances of the digital by applying our originally separate archaeology and art practices through one another's fields. Beyond this, we have experimented with some novel, combined, and metabolic art/archaeology practices. We employ the word *metabolic* to suggest that the way these practices combine is more than a simple intersection of practices. Metabolism involves the transfer or conversion of material, quite literally a stuff exchange. The conversion of matter, or

Stoffwechsel, was identified as early as the 1800s by Gottfried Semper (1803–1879) in his architectural writing:

> When an artistic motif undergoes any kind of material treatment, its original type will be modified; it will receive, so to speak, a specific colouring. The type is no longer in its primary stage of development but has undergone a more or less pronounced metamorphosis. If the motif undergoes a new change of material as a result of this secondary or even multiple transformation, the resulting new form will be a composite, one that expresses the primeval type and all the stages preceding the latest form. (Semper, 2004, p.250)

Thinking about the metabolism of vibrant matter (Bennett, 2010) in this way makes intangible things and tacit knowledge perceptible. Haraway also thinks through ideas about the organic transfer of matter when describing the ethics of collaboration as a form of composting. She argues that ‘staying with the trouble requires odd-kin; that is, we require each other in unexpected collaborations and combinations, in hot compost piles. We become-with each other or not at all.’ (Haraway, 2016, p.4)

One phenomenon that we have been exploring is the disturbing way in which archaeological and artistic artefacts metamorphose into unique stand-alone objects through that fallacious space described by Agamben (2019, p.IV) as ‘the mythical fixity of images[s]’, where artefact and art works are presented (de)contextualised within illusionary timeless, motionless, and apparently empty spaces. In these uncanny spaces, traces of both the object makers, their processes, and embodied practices of revealing and recording the object and the image makers—in short crucial paradata—are deliberately airbrushed away, masked, or cropped out. Such conventions have been common practice in archaeology since photography was first introduced into the discipline more than a century ago (e.g. Baird, 2019; Knight & McFadyen, 2019; McFadyen & Hicks, 2019; Thomas, 2019; Witmore, 2007). Derrida (1994) argued that the act of selection forecloses other alternative futures, but paradoxically these lost potential futures can haunt current and historical discourse and have real effects. The decision to not include the embodied worker and their deep skilful practices haunts countless images with their absence in the art and archaeology literature. Besides being ethically questionable, these absences create pedagogical chasms which deny students fuller, multimodal apprehension of the embodied practices required to produce the incomplete assemblage, as would-be students are deprived of much valuable tacit, or implicit, knowledge about how tasks of recording can themselves be usefully (re)presented. We argue that these diffractive images surface embodied knowledge pertaining to how practitioners perceive and communicate kinaesthetically. They emerge as potential pedagogic boundary objects that bring non-conscious and tacit knowledge into the explicit domain, hopefully to provide further pedagogical articulations to a broader set of learners/sharers for whom written or spoken accounts do not convey the full intent of the embodied practitioner (e.g. Derudas & Berggren, 2021). We want to make perceptible the image makers and some of the constraints and choices that affect the decisions made when (re)presenting artefacts or assemblages in images. In other words, we want to expose practitioner paradata that are normally excluded from, or deliberately disguised in, images. To this end, we adopt a diffractive transdisciplinary approach that also interlaces and entangles several standard cultural heritage recording methods to reveal some of these 'silent processes' of documentation (Huggett, 2020), the 'fingerprints' of researchers (Jones & Bunn, chapter 'Mapping Accessions to Repositories Data: A Case Study in Paradata' in this volume), and the subtly different 'paradata dustings' that different art and heritage practitioners produce (Buchanan & Huntsman, chapter 'Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites' in this volume).

Taken literally, diffraction describes the interference of light waves when they encounter an obstruction, and much has been talked of about the patterns that this interference creates by physicists and philosophers. Haraway introduced the concept of *diffraction* to contest the paradigm of *reflexivity* which 'mirrors the geometrical optics of reflection' (Barad, 2007, p.72). If the two-way reflective approach nurtures sameness, diffraction is remarkable for revealing the patterns of difference. Haraway harnessed this diffractive metaphor to discuss how better to account for the effects of the researcher on the experiment. Following Haraway, Barad, a feminist physicist, developed diffraction to much more than a metaphor. She argues that diffraction, actually and not just metaphorically, causes patterns that make a difference. 'Diffraction not only brings the reality of entanglements to light, it is itself an entangled phenomenon' (ibid., p.73).

Here, we apply Art/Archaeology diffraction filters to the concept of paradata and to their further layers of recursive introspection, supplements that we call *peridata*. We intend to reveal hidden paradata and peridata haunting the pixels of our so-called diffractive images (Dawson et al., 2022). After describing the object of study, we will define in more detail what we mean and intend by the terms 'Art/Archaeology approach', 'paradata', and 'peridata'. Then we will demonstrate how these para/peridata come about, non-consciously as practitioners' ecologies of attention constantly shift within a dynamically adjustable cognitive assemblage consisting of the artefact under investigation, the instruments of analysis, the changing position and settings of both instruments and practitioners, and the constant visual recalibrations of our knowledge that are intended to help 'share' our knowledge about the artefact. We will proceed, adopting a diffractive art/archaeology approach, by subversively applying several well-known cultural heritage recording and presentation techniques, recursively and unconventionally, to throw fresh light on embodied paradata and (re)present them as potentially very valuable pedagogical boundary objects. Rather than allowing artefacts and their paradata to subsist in inert, isolated, and sanitised vacuums, we will attempt to dislocate paradata away from a purely epistemological dimension and reposition them in an energetic, entangled, onto-epistemological nexus.

2 Introducing the Nessglyph

Our project centres on a remarkable carving, made on a block of red sandstone, found during the 2021 excavations of an Iron Age hillfort called Nesscliffe Hill Camp, Shropshire, UK (Hume & Jones, 1959; Lock & Reilly, 2019, 2020, 2021).

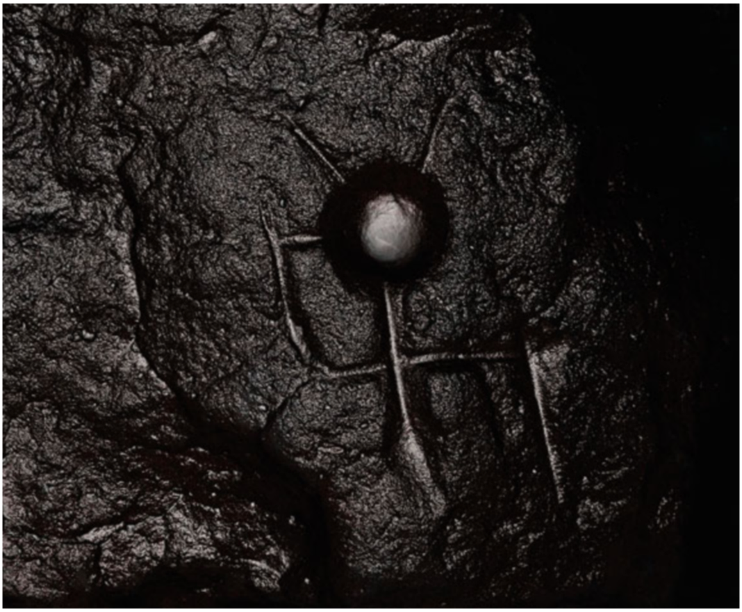


Fig. 1 The Nesscliffe Petroglyph, dubbed the *Nessglyph* (frame from an *interactive* RTI)

This pluritemporal petroglyphic assemblage known as the *Nessglyph* is a unique carved sub-triangular block of red sandstone (400 mm × 390 mm × 250 mm at its widest points). The Nessglyph is composite. Nevertheless, it *feels* as if it was composed, and it is certainly made by a sequence of overlapping marks, and therein lies its remarkability—an *image in the making* (Fig. 1).

The Nessglyph consists of a 'cup mark' that was later inscribed by a series of linear grooves. We examine its pluritemporal development, (re)composition, and interpretative possibilities in detail elsewhere (Reilly et al., 2023; Reilly & Lock, 2023). In this art/archaeology study, we are more concerned with looking at our own approach, our different ways of looking and knowing, such images in the making, and the creation of the paradata that we invest in, and attach to, the object. This image in the making has been defined by at least seven interventions affecting the biography and itinerary of the Nessglyph: initially the cup mark was ground out; then linear grooves were overscored; the stone (in the entrance) was buried; the stone was then disturbed unnoticed in a poorly controlled excavation; and once again reburied in backfill; in 2021 the stone was recovered, logged, and identified; the stone/Nessglyph is recorded using modern 3D photogrammetric and scanning devices. The last four archaeological interventions are likely to bear evidence of non-verbal paradata dustings and fingerprints. In the next section, we will begin to develop what we mean by non-verbal paradata and how it is fundamentally different from 'traditional' paradata. We start by laying out the case for apportioning greater weight and value to this previously overlooked source of embodied 'investigative documentation'.

3 Towards Alternative Embodied Perspectives on Paradata

The meaning, modes, and values of paradata are not homogeneous across disciplines (Andersson, Sköld, & Huvila, chapter 'An Introduction to Paradata' in this volume; Börjesson et al., 2022; Buchanan & Huntsman, chapter 'Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites' in this volume; Papadopoulos, chapter 'A Leap of Faith: Revisiting Paradata in 3D Scholarship' in this volume). In archaeology, the latent value of the concept was captured, but not recognised as such, in the rich alternative possibilities that 1980s 'data exploration loops' offered (Burridge et al., 1989; Reilly, 1989) within 'virtual archaeology' excavation sites (Reilly, 1989, 1991, 1992), in which practitioners of field archaeology could ask 'what might happen if we did things differently; what might we find otherwise?' Oddly, the term 'virtual archaeology', and subsequently 'virtual heritage', became synonymous with interpretive visualisations of reconstructed buildings. A growing number of these imaginative projects were delivered by expert modellers who were not trained archaeologists nor architects. Questions began to arise about the authority of the model makers. For example, how sure were they that this or that detail of the model was securely established? (e.g. Messemer, 2016; Miller & Richards, 1995; Opgenhaffen, 2021; Wittur, 2013). Over time, the need to augment these models with paradata pertaining to the veracity, degree of confidence, and possible alternative interpretations was formalised in the London Charter and Seville Principles (Bendicho, 2013; Bentkowska-Kafel et al., 2012; Börjesson et al., 2020; Opgenhaffen et al., 2021; Papadopoulos, chapter 'A Leap of Faith: Revisiting Paradata in 3D Scholarship' in this volume). The original focus on alternative perspectives and approaches in archaeological field and laboratory work was renamed 'digital creativity in archaeology' (Beale & Reilly, 2017a, 2017b). It is to fieldwork, laboratory, and studio work, that we now turn, to begin looking at embodied modes of decision-making (i.e. paradata) for recording and presenting artefacts and assemblages.

4 Techne, Poiesis, and Empeiria: 'Am I Doing This Right?'

Artefacts and assemblages do not exist in vacuums. They are not found or magically appear. They are produced. Removing soil and other materials by troweling, mattocking, or shoveling, exposing buried deposits, are central skills of the field archaeologist. Volunteer newcomers and novice students on training excavations frequently, and quite rightly, ask 'am I doing this right?' This is also a common refrain in art school studios as students navigate the entangled and shifting histories of *techne* (technique), *poiesis* (bringing something into being), and *empeiria* (practice without knowledge). Techne—the knowing how to make—has become a purely technical and technique-bound question associated more readily with craft and not art. Art is often more easily conveyed like empeiria as something that doesn't depend on rules and cannot be taught but is to be absorbed instead through non-verbal learning. Problems arise however from making assumptions that technique is somehow separate from art and that it has a lesser role in the development of an artwork. There is now an opportunity to rethink the role of techne in art, to consider technique as woven again into art making and to reconsider the distinctions between techne and empeiria (Elkins, 2001, p.103).

This apparently innocuous question—'am I doing this right?'—opens pedagogical issues across both disciplines, from the art student who is confused with how to think through making to the archaeology student when performing a seemingly simple mechanical action which is fraught with many imponderables (Pijpers, 2021). Can art students be taught? (Elkins, 2001). Similarly, how do you teach the archaeology student how to navigate through opaque deposits? Best practice in field archaeology is very relational. The soil matrix, weather, light conditions, and accessibility to the deposits all must be negotiated on a case-by-case basis. Certainly, an experienced excavator can usually determine if the novice is making a mess of matters, but it is not at all straightforward explaining how you adapt to the material circumstances immediately in front of you. Some students just seem to grasp how to explore deposits using handheld tools almost instinctively. Others struggle. How to articulate themselves and their tools does not come naturally to many students, who may then decide that fieldwork is not for them. Some, however, will persevere and gradually develop their technique and skill levels to learn to know through the tools they wield with their hands. We do not believe that troweling or mark-making more generally is necessarily an innate skill. It seems to us that the issue is really a pedagogical shortcoming and that perhaps embodied paradata could give practitioners another avenue to share their craft skills via another medium.

We are increasingly aware that movements, gestures, and rhythms of troweling, something that Tringham (2016) describes as 'hand-ballets' and Pijpers (2021) characterises as 'worlding with trowels', are fundamental to archaeological practice (see also Edgeworth, 2012; Wendrich, 2012) and can form rich sources of meaning-making and non-verbal decision-laden paradata (Gant & Reilly, 2018; Reilly et al., 2021). For instance, artist Stefan Gant's extended drawing studies of moving trowels in the hands of archaeologists at work showed skilful 'gamuts' of motion, with distinctive phrasing unique to individual diggers being deployed to probe and detect buried archaeological features. Initially, looking over the diggers' shoulder he traced their unique, deft, movements of the trowel's blade in pencil upon a sketchpad as their excavation proceeded. Complex 'meshworks' of normally unarticulated decision-making emerged and became available for study from the perspective of drawing theorists. These troweled gamuts form distinctive signatures (Gant & Reilly, 2018) of the otherwise anonymous diggers (Everill, 2009; Huvila, 2017). The start and end of each inscribed transect were initially read as a form of nonconscious paradata recording 'moments of tension' (Ingold, 2007, p.79) and 'moments of completion' (ibid., 81) at the 'trowel's edge' (Berggren & Hodder, 2003). However, while the overall gamut of movements can be discerned in these drawings, the sequence of marks made by both the digger and the mimetic transcriber were not always obvious. Digital video recordings were enlisted to address this issue. Gamuts of movement could now be (re)analysed frame by frame and (re)materialised as 3D extended drawings. Each trowel mark in an extended engagement with arbitrarily selected archaeological features was reverse engineered by cutting card strips to the appropriate length and orientation of the trowel's indents and stacking them in reverse order. After applying the well-known Cultural Heritage Imaging technique known as Reflectance Transformation Imaging, or RTI (CHI, n.d.), to this 3D extended drawing of troweling marks, the morphing temporal diffraction shadows (Callery et al., 2022) we observed under several RTI filters impressed upon us that these truncated, interweaving gamuts did not constitute meshworks. Rather, these troweling gamuts might be better conceived of as sinuous 'knots' (Ingold, 2015) of meaningful motion in which the decision about where and how to apply the trowel occurs, non-consciously, during the excavator's expressive world-making looping of the tool before and after the trowel's blade scored the ground. Some of this meaning-making action is also detectable in the accompanying acoustic registers. Recording the archaeologists' mattocking, shoveling, and troweling different materials and displaying this activity as sonography gave new 'voice' to these workers, and their skilful use of their tools, as the contexts and artefacts they encountered emerged. Distinctive 'sonic stratigraphies' could be detected in the soundtrack of these unchoreographed 'hand ballets' of discovery when they are presented as acoustic and visual paradata (Gant & Reilly, 2018; Reilly et al., 2021). We have exposed significant underlying patterns of expression through these initial observations, but the precise articulations and subtle gestures of the hands of the working archaeologists in these dense loops of meaning-making are still partially withdrawn from these extended drawings made in the field. Perhaps we can get more traction on subsequent gestures of meaning-making and authorship (i.e. visual, haptic, and acoustic paradata) made in the better controlled environment of the finds hut, post-excavation laboratories, and artist studios (e.g. Min et al., 2020).

Jones and Smith (2017) foregrounded the performative nature of producing Reflective Transformation Images (RTIs). Dawson and Reilly (2019) then realised that RTIs contained 'inadvertent images' (Geimer, 2018) recording autographic traces of the RTIs in their making. All RTIs naturally operate as 'metapictures' (Mitchell, 1994, 2004) embodying a self-referential quality that triggers a meta-level discursive opportunity to consider what, when, where, and how, this form of technical image operates. Painted in light on the crucial RTI sphere is a significant amount of the set, the choreography, and the prompts surrounding these RTI 'performances' captured in a time-lapse sequence as each image is produced.

The artefact or assemblage intended to be recorded, the DSLR device, the strobe, the reflective sphere, and the image maker, already entangled, are further entangled by residual traces of light. The reflections caught on the surface of the RTI sphere can be thought of as the spontaneous and co-authored signature of the total assemblage (Fig. 2). They are also another form of auto-archived visual paradata, recording the circumstances, environment, relative position, poses, gestures, and the condition of all the actants and their intra-actions in this emerging 'assemblage of

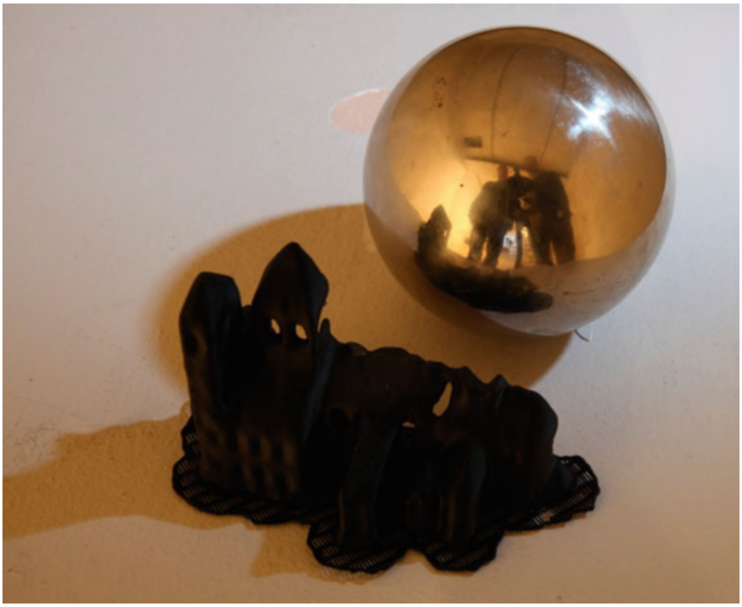


Fig. 2 Auto archiving paradata via a RTI metapicture

practice' (Antczak & Beaudry, 2019) or 'cognitive assemblage' (Hayles, 2017) as it unfolds from image to image. RTI can, quite literally, shine a light on what the image maker is most concerned to bring forth from the artefact or assemblage (Callery et al., 2022).

5 Shining a Light on the Image Makers

We attempt to extend these insights concerning non-verbal paradata through four techniques. Three are familiar to archaeology and cultural heritage: Highlight-Reflectance Transformation Imaging (H-RTI), Structure from Motion photogrammetry (SfM), and 3D Structured Light scanning (SL). The fourth is Fused Filament Deposition (FFD), the technology of 3D printing, an emerging area in the cultural heritage sector as a supplementary tool for tactile engagement (e.g. Reilly, 2015; Reilly & Dawson 2021; Reilly et al., 2016). All these processes are additive. In RTI the subject under documentation and the digital SLR (DSLR) are held static and a series of images of the subject are taken as strong directional lights are moved around the subject, eventually forming a dome of lighting positions, to bring out details. The outcome is a synthetic model of the recorded object that can be interactively relit, and its surface properties redefined to allow, for example, specular enhancement (Fig. 1). In the case of SL scanning, bands of projected light haptically stroke the object of study in order to capture its surface geometry. In

SfM, digital photographs taken from multiple different but overlapping viewpoints are also processed and assembled into synthetic images or three-dimensional models that can be rotated, panned, and zoomed, interactively. With FFD printing the digital model is cut into tiny slices, turned into a sequence of silhouettes to be printed in material layers.

We apply these technologies in (un)familiar ways to test the documentation processes that we hope will further the decipherment of the Nessglyph. We apply standard, and experiment with 'dirty', versions of H-RTI, SfM, and 3D SL scanning. Each of these three techniques in their pure form has associated best practices, all involving movement to take multiple overlapping impressions of the subject under study (Historic England, 2018a, 2018b). Images conforming to the best practices of each of these techniques produce a genre that tend to look very similar to one another, which is hardly surprising as they are composed in conventionalised poses and executed under constrained parameters. This means that, in practice, many of the decisions needed to record an artefact are taken away from the operator 'in charge' and are instead prescribed or delegated to the nonconscious cognition of algorithms and sensors concealed within technological or cognitive assemblages (Hayles, 2017; Huggett, 2017) that the operators handle. Flusser (2011) argues that the role of the operator in the making of 'technical images' has been reduced to that of a mere 'functionary', someone who just must point the instrument in the appropriate way and press a button. Cubitt (2014, 270) goes further and claims that operators of these instruments are 'enslaved' to the technology. Crucially, in the context of this discussion, the paradata (i.e. the decision-making criteria and constraining factors that had to be overcome) are now rendered silent and invisible in their black boxes. Ironically, their disappearance draws attention to their whereabouts and prompts further questions regarding their epistemic status in our knowledge frameworks. In other words, the provenance of the selected paradata to record is also called into question. Why were these paradata chosen as being more important than other options in the first place? These emerging layers of introspection—that is paradata about the selection and use of the paradata that are articulated—are called 'peridata' (Gant & Reilly, 2018).

Significant knowledge is certainly buried within these technical assemblages. However, we will argue that the performative images remain images in the making and therefore retain opportunities to break free of the shackles of the functionary and expose some of the 'silences' in the data, metadata, paradata, and peridata (Huggett, 2020; Ortolja-Baird & Nyhan, 2021)—those nonconscious and tacit aspects in our digital imaging knowledge making—through visual forms of diffractive analysis. Instead of attempting to consciously document the decision processes that culminated in our models, we create digital skeuomorphs of the Nessglyph in order to identify and explore, diffractively, non-conscious paradata, and non-verbal gestural peridata, embedded in the image datasets that were derived from the artefact. Skeuomorphs are objects or features produced in one medium that mimic, or cite, the processes of making, or the inherent properties, of a similar-*appearing* prototype made using another medium with different properties (e.g. a ceramic pot moulded to look like a woven wicker basket). Our approach is to record one documentation

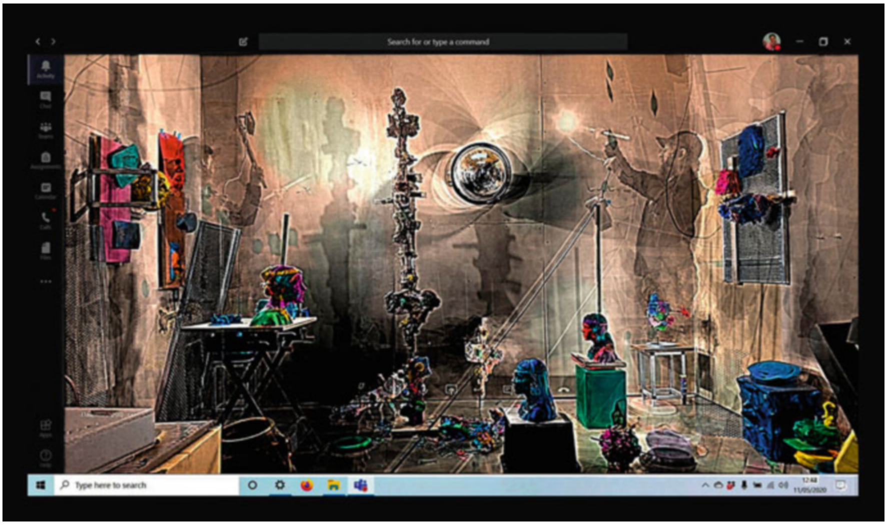


Fig. 3 Remote DiRTI scene of Ian Dawson's Plastic Studio (2022)

process through the lens of an ontologically different technological assemblage, including the embodied movements of the operators. So, for example, we will trace the making of a SL 3D scan through the uber-cyborgic gaze of a *dirty-RTI* (DiRTI[1]) session (Dawson, 2020). Our focus is not the Nessglyph per se, nor its making, nor its interpretation. Here, we (re)consider our documentation processes diffractively through a form of visual introspection that enables us to critique our own embodied and mediated interactions with the artefact. Through this subversive art/archaeology manoeuvre, we endeavour to record how *we* document the Nessglyph, and extract visual traces of both the embodied and cyborgic decision-making processes (i.e. paradata) in their making. Crucially, we deliberately put ourselves in the frame to haunt the pixels we render (Fig. 3).

In short, we offer a non-standard, transdisciplinary, and diffractive analysis of embodied art/archaeology gestural approaches to recording this carving and, specifically, what we call auto-archived non-verbal paradata and potential convolved peridata (i.e. paradata about the paradata and other related metainformation scenarios (Huggett, 2023; Martin-Rodilla & Gonzalez-Perez, 2019)).

The 'technical images' we develop, using these (non)standard documentary techniques, expose the effects of tell-tale gestures and actions, and allow us to recognise and reconsider significant non-conscious decisions (i.e. spontaneous paradata) made by practitioners within the overall cognitive assemblages through

[1] We claim no credit for either the term or the acronym. Eleni Kotoula first uttered, in shock, the term 'dirty-RTI' (Dawson, 2020, p.62), and we have Wout Dillen to thank for this neat abbreviation (i.e. DiRTI).

which these digital datasets are collected, at what density, and from what specific viewpoints, and what is lost in this blending. Each of the three standard recording methods employed produces interactive images, which means that part of the process of analysing these data (and, consequently, the paradata) can be outsourced by the maker to diffractive viewers when—most significantly here—the paradata about decisions on how to interact with these (para)data can become untethered. Nevertheless, the collected data (i.e. images) and assembly files may still be used to reveal careful, but perhaps unconscious or non-conscious, decisions about what matters to the practitioner in the making of the model. For example, in both RTI and SfM imaging every viewpoint is embedded and (re)displayable in the finished model. We can therefore begin to reverse engineer the practitioner's focus of attention and, to a degree, their 'intent' by showing, for instance, which parts of the object under investigation bear more scrutiny than others. Equally, we can determine some of the constraints imposed by the process of data collection. For example, we experiment with a subversive art/archaeology version of Reflectance Transformation Imaging (RTI), called *dirty-RTI* (DiRTI), to document, recursively, both a structured light 3D scanning and RTI-recording session. During the DiRTI sessions we capture non-conscious decision-making gestures and movements of the practitioners made in the *mise en abyme* occupying the space between a panopticon-like digital camera and the petroglyph.

Later, we extract geometric and other surface data, such as colour and surface normal from our image datasets and reuse them to create 3D prints of the artefact. During this metamorphic process much of the previously auto-archived gestural paradata outlined above is severed, and a whole new layer of physical and digital (phygital) paradata physically congeals in these 3D material (re)renderings of earlier image datasets. Decisions in the workflow—which include, for instance, digital coring and slicing—leading to printing components in a particular orientation and then their physical assembly all become evident in the object, giving these latest skeuomorphs a set of interactive parameters of their own. In some instances, in an odd inversion of properties, such as specular enhancement, these new objects become analogous to RTI when held and rotated in one's hand.

6 Paradata in Motion

The Nessglyph has been subject to a series of documentation processes familiar to archaeologists and art historians. Figure 4, for example, is a rough sketch made by Reilly in his scuffed and blotted day journal. These reflexive notes illustrate the usual sorts of metadata, such as dimensions, annotating sketches of new finds, and developing interpretation of features as they emerge. Field notebooks are paradata par excellence. Unfortunately, in terms of being knowledge containers for knowledge management they are problematic. They are highly idiosyncratic, unstructured, and, as apparent in Fig. 4, often badly treated by wear and weather. As Huggett (2020) observes, the journals of the fieldwork team do not feature in the final hegemonic reports that get published under the names of the authoritative

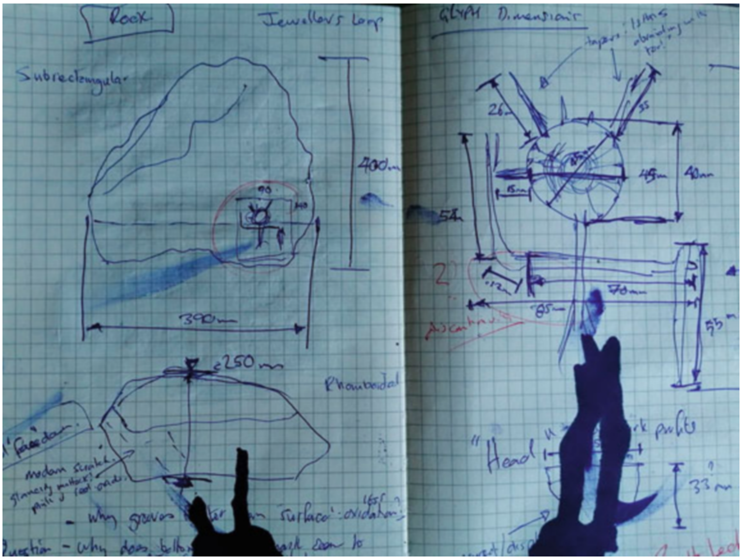


Fig. 4 Typical archaeological paradata: annotated, weathered, and blotted sketch of the Nessglyph in a day journal

principal investigators. They may not even reach the physical archive. With some exceptions, these thoughtful notes will go mostly unread. Although internalised by the individual practitioner, this knowledge is not widely shared and so the valuable written and drawn insights and learning points become lost opportunities, that is orphaned paradata.

Far more likely to find their way into a knowledge base and, perhaps, the hegemonic site report, are the carefully constructed, and *processed*, images (Morgan & Wright, 2018). Compare how the carved Nessglyph is presented in Fig. 1 compared to Fig. 5. The incised sandstone block is showcased conventionally in Fig. 5. In Fig. 1 the Nessglyph is presented in an apparently static, empty, featureless, and distraction-free space (except for the bottle holding the reflective sphere, which also operates as a scale). However, this illustration disguises many important aspects of the Nessglyph, not least the circumstances and environments surrounding its discovery and its recording, and, crucially, it forecloses other potential interpretations (Derrida, 1994; McFadyen & Hicks, 2019; Thomas, 2019). To begin with, despite appearances, the exhibit in Fig. 1 is not a photograph. It is a frame from an interactive polynomial texture map (.PTM) produced using the technique called Highlight-Reflectance Transformation Imaging (H-RTI) (Historic England, 2018b). Similarly, Figs. 2, 3, and 5 are synthetic views derived from a series of digitally skeuomorphic photographs, using skeuomorphic photographic technology and in the genealogy of analogue photographic tropes (Jones & Díaz-Guardamino, 2019; Taylor & Dell'Unto, 2021), but updated to create something Flusser (2011) characterises as a 'technical image'. Like all technical images, they

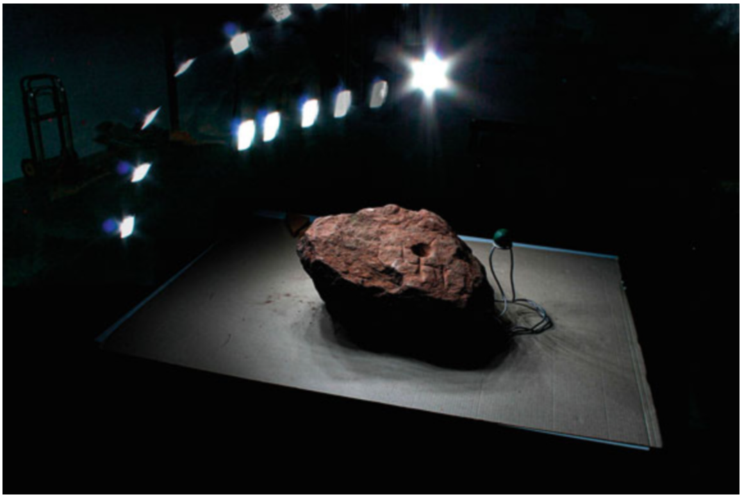


Fig. 5 Frame from an H-RTI session recording the Nessglyph

have more in common with spreadsheets than pictures (May, 2019). As Galloway (2011, p.88) emphasises, data in their purest form exist as numbers and therefore their 'primary mode of existence is not a visual one'. To create this technical image, data were 'assigned visual characteristics and converted, or rather translated, into what we humans recognise as photographs' (Zylinska, 2017, p.26) However, as Rubenstein (2020, p.4) points out, the same data could equally be 'output not as an image file but as a sound file, a text file, as a string of numbers, or it could be left unprocessed.' So, whereas the material artefact is a heavy carved red sandstone block that is difficult to handle, the digital skeuomorph (re)presented in Fig. 1 is a dimensionally elastic, weightless, surface apparition. The rendered surface of this simulacrum has no thickness and envelopes nothing, but is susceptible to panning, zooming, and relighting. The seemingly empty, inert space surrounding the digital simulacrum is equally skeuomorphic and just as deceptive. This apparent void is defined in the same code and data definitions as the featured artefact, and potentially contains the memory of light, shot through with movements that are traces of decisions and choices in motion, which flooded the recording session as it progressed. We will reveal some of these energetic, non-verbal, paradata haunting the pixels, and their underlying data that were intended to (re)present the Nessglyph.

We have been on the trail of non-verbal paradata for a while now. During the pandemic, forced to collaborate at a distance, we began to consider how the widely used heritage recording technique of RTI could throw light on itself as both a technical assemblage and a process. After seeing ourselves so often within *Teams* and *Zoom* sessions it was perhaps inevitable that we would try to find ways of breaking away from the tyranny of technology (pace Flusser (2011) and Cubitt (2014)) and subvert, bend, or break some rules. We decided that instead of hiding ourselves behind the image-making process we would deliberately situate ourselves within the frame of the camera in order that we could, later on, observe our embodied decisions about where and how to position the all-important and meaning-making light source in each successive frame of the RTI (Dawson, 2020; Reilly et al., 2021).

The interactive images that began to emerge from these sessions were startling (e.g. Figs. 1 and 5). The 'empty' studio space was now filled with metabolic energy and movement (Fig. 5). The studio space also became plastic and metamorphic. The end user experience of this dirty form of RTI file through the RTIViewer is transformed as ghostly apparitions come in and out of view as the user interacts with the RTI. This is possible because RTIs are miniature quantum-like universes where the linear arrow of time does not apply. What was previously 'before-and-after' now stumble around one another through quantum superpositions of pixel properties (Callery et al., 2022). We realised that through these images we could watch our decision-making (un)folding, that paradata could now haunt the pixels of the interactive image, and, perhaps most interesting of all, that we had stumbled upon a back door for the functionary—a way into, and back out of, the black box of RTI.

We wondered if we could extend this insight into investigating the other two standard archaeological computational photography methods that we had already deployed on the Nessglyph. Could we diffract similar kinds of embodied paradata that are interstitial to Structured Light (SL) scanning through a DiRTI process to reconsider our unthought embodied decisions—the indirect visual haptic strokes of the scanner at a distance—as diffracted peridata?

As we have already mentioned, in RTI the ostensible subject of study—accompanied by a reflective sphere and the imaging device (DSLR)—remains static in the recording session. All the action takes place between each frame being taken when the strobe or other light source is moved around the set. Highlights on the reflective sphere enable the RTI software to simulate all the various lighting angles required to interactively relight the model. By contrast, working with the Artec hand scanners available to us, the image maker has to constantly negotiate with the object. The lasers record thousands of readings a second in waves of cold flickering white light. It occurred to us that we could use those flickers as the directional light source for a DiRTI session on the Nessglyph (i.e. a hybrid or interlaced SL and DiRTI recording session run concurrently). Experimentally, we discovered that by adjusting the exposure for each RTI image to about 3 s could produce stunning hybrid forms of extended drawing to help us reconsider the gestures made by the operator.

Figure 7 is a single frame from one such interactive SL/DIRTI hybrid recording session. Here the graceful knots of movement that eluded us in the RTI of the stacked cardboard trowel marks produced with Stefan Gant (Gant & Reilly, 2018) snake and twist in the form of mottled ribbons of light around the Nessglyph when the compiled composite dirty-RTI is inspected through the RTIViewer. These snakeskin-like ribbons of light are the traces of remarkable unsighted hand ballets.

While the scanner operator must look away from the Nessglyph and concentrate on the screen informing the functionary about which parts of the stone are being

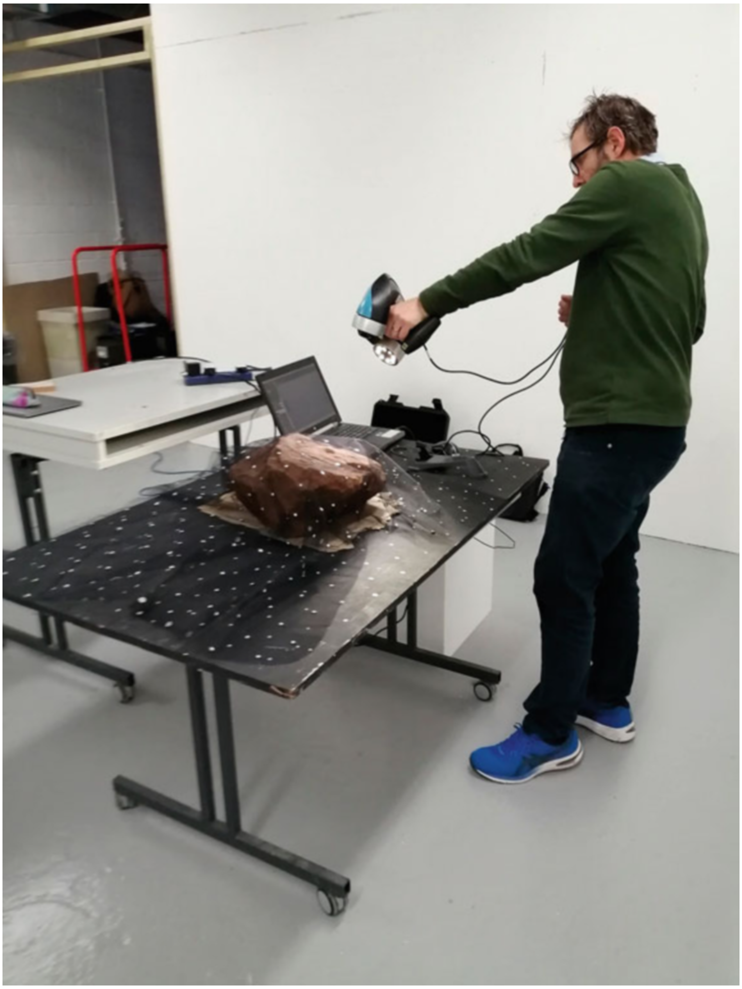


Fig. 6 SL Nessglyph scanning session note operator looks at laptop

rendered by the flickering lasers (Fig. 6), the slower eyes of the RTI trace subtle and delicate twists, turns, and rolls of the hand-operated scanner. Unseen by the operator, the stone is being stroked by strobing light at a constantly adjusted hovering distance from the surface of the stone to expose every mark incised on the red sandstone block in interactive detail. Decisions about where to point and wave the lasers to bring out important details are made haptically through fine adjustments of proprioception in the hand, wrist, and forearm of the operator. These gliding, handheld, scanning gestures bear an uncanny resemblance to the actions of the troweling archaeologist trying to envisage the form of buried artefacts and

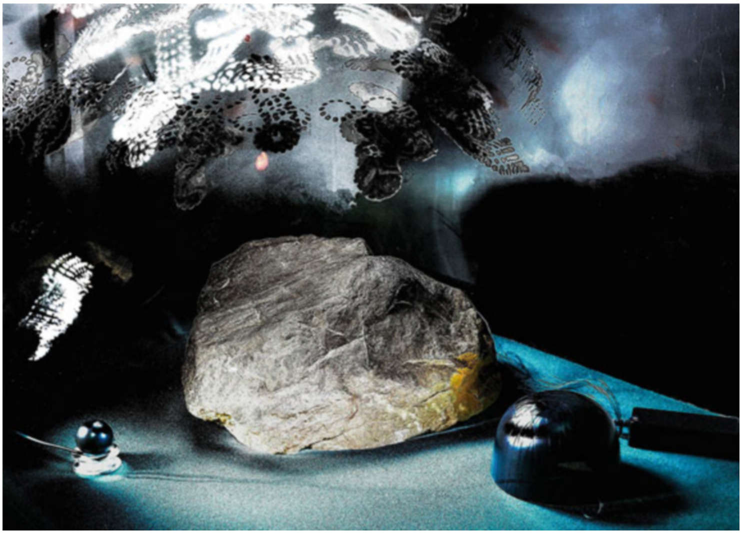


Fig. 7 Diffracting structured light scanning through dirty-RTI

features hidden from direct view. These SL paradata now figure in the RTI paradata (Fig. 7).

7 Phygitally Dislocated Paradata

As Mitchell (2003, p.3) declared two decades ago in *Me*++, the separation of bits (the elementary unit of information) and atoms (the elementary unit of matter) is over. With increasing frequency, events in physical domains reflect events in virtual domains. Phygital information can, for example, direct the movement of the printer nozzle of a 3D printer to produce another skeuomorphic iteration of the Nessglyph. Now the movements and gestures (i.e. embodied and gestural paradata) of the original glyph maker(s) and the art/archaeological investigators are interlaced with the post-human cyborgic gestures of an additive manufacturing fabrication process. For example, the physical (re)presentations of the Nessglyph shown in Figs. 8 and 9, produced by slicing the digital prototype into a sequence of silhouettes, allow for a new phygital skeuomorph of the original artefact to be 3D printed in layers.

To accomplish this, the SL scans were converted into three packages of data: an OBJ file, which defines its geometry as a sequence of code that lists the xyz coordinates of the vertices of the object, plus a Material Template Library (MTL) file providing the lighting information, with a texture map (often saved in a JPG or TIFF format) wrapped over the OBJ code. These were converted into an STL file in order to 3D print the model. The STL file, developed by 3D Systems as part of their development work with the 3D Stereolithographic process, is another sequence

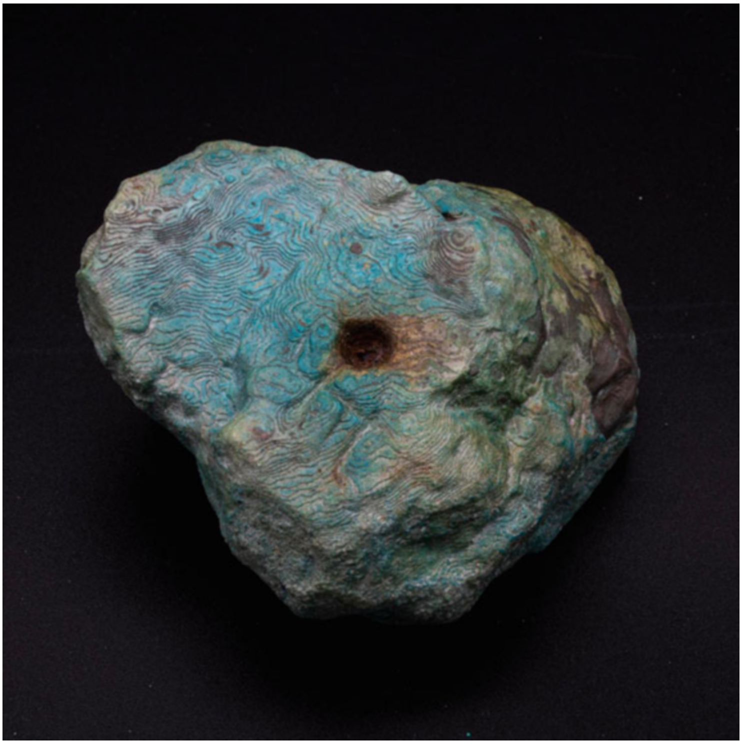


Fig. 8 3D print derived from a structured light scan responding to the idea that the Nessglyph might have been carved into the stone with a bronze tool

of code which enables a model built from interconnecting triangles to be easily sliced horizontally in preparation for the 3D print process. On the 3D printer these horizontal layers are printed in a continuous sequence. A PRUSA i3 MK3 printer was used to print the Nessglyph firstly on a reduced scale with a plastic/bronze amalgamated filament (Fig. 8). This object changed colour *like* a conventional bronze; its green patina referencing its metabolism as it slowly oxidised. The action of the 3D printer is captivating, like watching a stylus on a vinyl record. There are multiple enchantments that the machine gestures towards. The 3D print we hold in our hands is a souvenir of the myth of the pure potentiality of the formless, and the capacity of the phygital to morph, recompose, and reformat. The 3D print as a souvenir performs the role of catalyst for a new narrative. Once extracted from the 3D print bed, the 3D printed Nessglyph substitutes the context of origin of the glyph with the second-hand experience of the handler. We are now in possession of a metonymic object. Such objects are but samples of a now past experience, and by their very own impoverished translation, the partiality of the souvenir is laid bare. These copies are allusions, not models. They function best alongside a loosely attached supplementary narrative discourse that creates a new myth with regard to

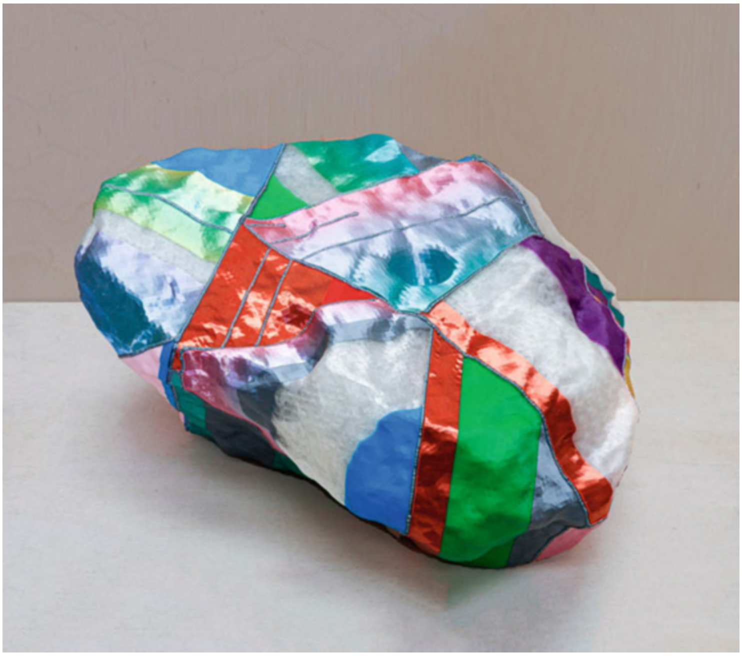


Fig. 9 Performative mattering: Nessglyph (re)presented using the wholly different set of gestures made by the plastic artist and a 3D printer

their origin (see Stewart, 1984, p.135). In other words, these souvenirs foster the need for further (mythical) paradata schema.

During this phygital metabolic exchange, the concepts of techne, poiesis, and empeiria have all been infused into a material amalgam that replaces the paradata associated with the initial digital model in a dazzling rendering of completely new, cyborgically expressed, gestures implemented in colourful plastics. In this metabolic transformation, links to any pre-existing paradata are severed. Paradoxically, these new mythical plastic instantiations of the Nessglyph amplify the initial inquiry: ‘what am I recording or (re)presenting?’ The outcome of our experiments it seems is that epistemological and ontological concerns are now entangled (Fig. 9).

8 Summary, Discussion, and Conclusions

We have argued that attempts by galleries, libraries, archives, and museums to present artefacts within empty spaces are misleading and divisive. These exhibits and accessions hide the considered, embodied, work and decisions—in other words the layers of paradata (peridata)—made by many largely invisible people, applying their considerable skill sets, knowledge, and experience, by framing and presenting the artefact in a disembodied featureless vacuum.

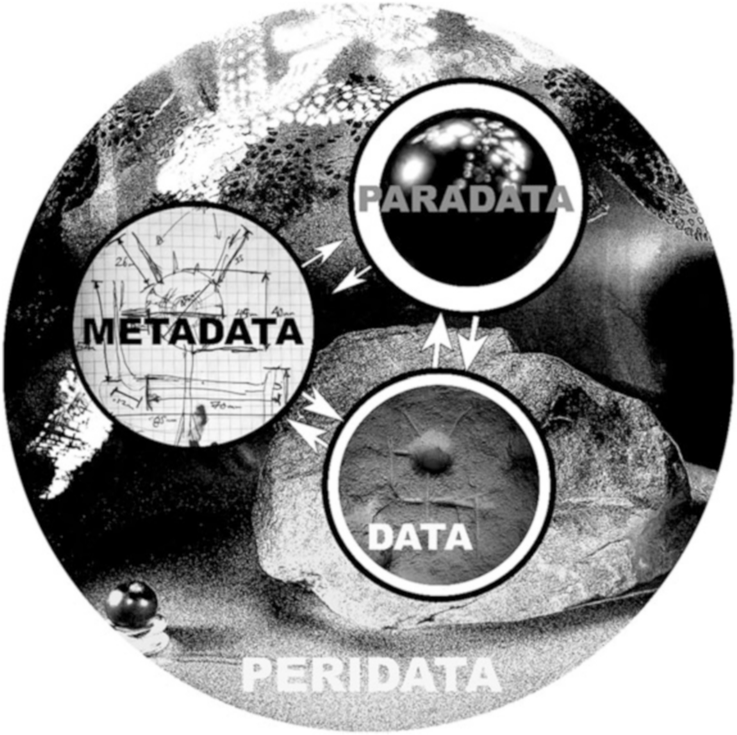


Fig. 10 Peridata: layers of (data)introversion. By choosing specific data, metadata, paradata to be recorded, we invite further layers of introspection: why were they chosen? (ad infinitum)

In contrast, by diffracting several imaging processes through one another, we have tried to help anonymous field archaeologists and artist assistants, and other enslaved technological functionaries, escape the shackles of automated image-making apparatus and the requirement for them to produce images that correspond to certain general conventions and configurations. We have also attempted to break away from conventional strategies of teaching embodied practices that rely on verbal instruction for learning. Our unconventional DiRTI approach has enabled many layers of recursion through Structured Light scanning and Structure from Motion photogrammetry information infrastructures and, admittedly, the images we create take considerable effort to unpick. At a top level, we have compiled ourselves and our meaning-making processes, recursively, into these DiRTIs of the Nessglyph (Fig. 3). The blended RTI sphere (Fig. 10, top right) highlights the diffractive shadows we cast on the data underpinning this RTI. Each flash of the strobe creates a metapicture which, when assembled, reveals the many different positions that we and our apparatus adopted during the shoots (Dawson & Reilly, 2019). What becomes clear is that the apparently inert, empty, space surrounding the object under study is an intersubjective space full of energy, light, movements, gestures, and equipment (Figs. 2, 5 and 7). These interactive intersubjective spaces are laden with meaning-mattering decisions and adjustments. We can think of DiRTIs as a form of 'autographic' image (Offenhuber, 2020). That is to say that it contains 'a trace of the process itself: it retains some interpretive authority, and it is taken as a product of the phenomenon at its face value' (Likavčan & Heinicker, 2021, p.212).

Autographic images, which include timelapses, live streams from space, and the RTI images that we have practised here, become phantom operational images (Farocki, 2004) as human labour is joined with the labour of computational algorithms. Autographic images are sensitive to the socio-material context (Fenwick et al., 2012; Pelizza, 2021) of their contrasting 'information infrastructures' (Huvila, 2019) and, just as in our hybrid DiRTIs, they reveal material phenomena as visible traces which draw the viewer's attention back to the intimate ways in which the previously opaque reality of the process of imaging can now unveil itself in the productive alignments between its human and non-human elements. These 'images in the making' contain things in motion, involving conscious and non-conscious processes of assembling and reassembling, and reimaginations of the world, in other words paradata. As Back Danielsson and Jones (2020, p.4, original emphasis) point out, 'if we understand *imaging* as a process of assemblage making, subsequent processes of viewing and intra-action are also components of the continuous process of *imaging*.' Put another way, these images can be understood as assemblages of ongoing processes and ongoing paradata in the making.

Remembering that all RTIs naturally operate as 'metapictures', the DiRTIs presented in this chapter are not simply epistemological models. The unconventional cognitive assemblages we have created also allow us to observe the observed and the observers. The images are not intended merely to serve as illustrations to a commentary on decision-making in practice-based disciplines like archaeology and fine art. They directly picture layers of paradata in the making (i.e. peridata). Figure 10 summarises this recursive process: the data of interest are selected and then assigned attributes (metadata); since both data and metadata are chosen there is a need in some intellectual quarters to justify how these decisions were arrived at (paradata). The blended RTI highlights are a form of visual paradata recording how and in what order the data were imaged. This blended DiRTI highlight image is a hauntology in which the black areas represent lost futures of analysis in this model. In these dirty sessions the highlight detection algorithm was often unable to disambiguate the true highlight of the source. They had to be manually edited. Unfortunately, the finished blended highlight file is not updated by the software with these manual corrections and so this summary image appears to give a much lower density and coverage of light source positions. However, the decision to record those paradata and their attributes remains just another layer of introspection. Put simply, one person's metadata or paradata may be another's data, requiring further introspection. These layers of introspection—we have called peridata—could, in principle, be never-ending.

The case for non-verbal paradata in art/archaeology as business-as-usual is therefore problematic since the reasons for elevating certain decisions to paradata status also need to be analysed and explicated (recursively via peridata). Problematic but not necessarily futile (contra Reilly et al., 2021). In this study, we have diffracted embodied paradata haunting several different forms of ‘undigital images’ (Zylinska, 2021) that were created and modified by archaeologists, artists, and cultural heritage workers more generally, to reveal spectral pedagogical boundary objects that enable us to share and exchange non-verbal, non-conscious, embodied, multimodal, decision-making processes across disciplinary boundaries and domains, and across different learning levels (apprentice, journeyman, master) and times. In their physical translations earlier paradata became untethered as new paradata of performative mattering reinstated the Nessglyph in novel dazzling plastic forms. In this admittedly subversive art/archaeology context, diffracting embodied paradata allows us to interlace different registers of techne, poiesis, and empeiria which then enables us to expose important points of difference and start reconsidering some of those elusive layers of tacit learning and teaching that underpin the development of skilful meaning-making embodied practices in field archaeology and art. In the process we have shifted the role of paradata from the reflexive epistemic considerations contained in, for example, traditional field notebooks to a diffractive position in which a relational ontology has emerged that can no longer be categorically separated from epistemological processes. Our entangled computational photographic and 3D printing methods of engagement are simultaneously generating radically immanent but relational new worldings. In these art/archaeology material-discursive experimental and metabolic entanglements, ontology and epistemology have become intra-laced, implying that both paradata and peridata have opened a new onto-epistemological dimension requiring (re)theorising.

Acknowledgements The fieldwork at Nesscliffe Camp hillfort has been funded by the Shropshire Archaeological and Historical Society (Pagett Fund), the Prehistoric Society (James Dyer Prize), the Society of Antiquaries of London, and the Robert Kiln Trust. Thanks also to Gary Lock (Co-director of fieldwork), the excellent volunteer team, plus Andy Wigley and Shaun Burkey (Shropshire Council), and Bill Klemperer (Historic England) for their guidance and unstinting support for the broader Nesscliffe Hillfort project.

References

Agamben, G. (2019). 11. Notes on gesture. In C. Kul-Want (Ed.), *Philosophers on film from Bergson to Badiou: A critical reader* (pp. 208–217). Columbia University Press. https://doi.org/10.7312/kul-17602-013

Antczak, K., & Beaudry, M. (2019). Assemblages of practice. A conceptual framework for exploring human–thing relations in archaeology. *Archaeological Dialogues, 26*(2), 87–110. https://doi.org/10.1017/S1380203819000205

Back Danielsson, I. M., & Jones, A. M. (2020). *Images in the making. Art, process, archaeology*. Manchester University Press. https://doi.org/10.7765/9781526142856

Bailey, D. (2017). Disarticulate—Repurpose—Disrupt: Art/archaeology. *Cambridge Archaeological Journal, 27*(4), 691–701. https://doi.org/10.1017/S0959774317000713

Bailey, D. (2018). Art/archaeology: What value artistic-archaeological collaboration? *Journal of Contemporary Archaeology, 4*(2), 246–256. https://doi.org/10.1558/jca.34116

Baird, J. A. (2019). Exposing archaeology: Beauty, time, and mistaken images. In L. McFadyen & D. Hicks (Eds.), *Archaeology and photography: Time, objectivity and archive*. Bloomsbury Visual Arts. ISBN 9781003103325.

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Beale, G., & Reilly, P. (2017a). Digital practice as meaning making in archaeology. *Internet Archaeology, 44*. https://doi.org/10.11141/ia.44.13

Beale, G., & Reilly, P. (2017b). After virtual archaeology: Rethinking archaeological approaches to the adoption of digital technology. *Internet Archaeology, 44*. https://doi.org/10.11141/ia.44.1

Bendicho, V. M. L. M. (2013). International guidelines for virtual archaeology: The Seville principles. In C. Corsi, B. Slapšak, & F. Vermeulen (Eds.), *Good practice in archaeological diagnostics. Natural science in archaeology*. Springer. https://doi.org/10.1007/978-3-319-01784-6_16

Bennett, J. (2010). *Vibrant matter: A political ecology of things*. Duke University Press.

Bentkowska-Kafel, A., Denard, H., & Baker, D. (2012). *Paradata and transparency in virtual heritage*. Ashgate.

Berggren, Å., & Hodder, I. (2003). *At the trowel's edge: an introduction to reflexive field practice in archaeology*. Westview Press.

Börjesson, L., Sköld, O., Friberg, Z., Löwenborg, D., Pálsson, G., & Huvila, I. (2022). Repurposing excavation database content as paradata: An Explorative analysis of paradata identification challenges and opportunities. *KULA: Knowledge Creation, Dissemination, and Preservation Studies, 6*(3).

Börjesson, L., Sköld, O., & Huvila, I. (2020). Paradata in documentation standards and recommendations for digital archaeological visualisations. *Digital Culture & Society, 6*(2), 191–220. https://doi.org/10.14361/dcs-2020-0210

Burridge, J. M., et al. (1989). The WINSOM solid modeller and its application to data visualization. *IBM Systems Journal, 28*(4), 548–568.

Callery, S., Dawson, I., & Reilly, P. (2022). Temporal ripples in art/archaeology images. In I. Dawson, A. M. Jones, L. Minkin, & P. Reilly (Eds.), *Diffracting digital images. Art, archaeology and cultural heritage* (pp. 97–119). Routledge. https://doi.org/10.4324/9781003042129-7

CHI. (n.d.). *Reflectance transformation imaging*. Cultural Heritage Imaging. Available: http://culturalheritageimaging.org/Technologies/RTI/.

Cubitt, S. (2014). *The practice of light: A genealogy of visual technologies from prints to pixels*. The MIT Press.

Dawson, I. (2020). Dirty RTI. In I. M. Back Danielsson & A. M. Jones (Eds.), *Images in the making. Art, process, archaeology* (pp. 51–64). Manchester University Press.

Dawson, I., Jones, A.M., Minkin, L., & Reilly, P. (2022). What is a diffractive image? In I. Dawson, A.M. Jones, L. Minkin, & P. Reilly (Eds.) *Diffractive digital images. Archaeology, art practice and cultural heritage* (pp. 1–14). Routledge.

Dawson, I., & Reilly, P. (2019). Messy assemblages, residuality and recursion within a Phygital Nexus. *Epoiesen*. https://doi.org/10.22215/epoiesen/2019.4

Derrida, J. (1994). *Specters of Marx: The state of the debt, the work of mourning and the new international* (Trans. P. Kamuf). Routledge.

Derudas, P., & Berggren, Å. (2021). Expanding field-archaeology education: The integration of 3D technology into archaeological training. *Open Archaeology, 7*(1), 556–573. https://doi.org/10.1515/opar-2020-0146

Edgeworth, M. (2012). Follow the cut, follow the rhythm, follow the material. *Norwegian Archaeological Review, 45*(1), 76–92. https://doi.org/10.1080/00293652.2012.669995

Elkins, J. (2001). *Why art cannot be taught (A handbook for students)*. University of Illinois Press.

Everill, P. (2009). *The invisible diggers: A study of British commercial archaeology* (2nd rev ed.). Oxbow Books.

Farocki, H. (2004). Phantom Images. *Public*, (29). Retrieved from https://public.journals.yorku.ca/index.php/public/article/view/30354

Fenwick, T., Nerland, M., & Jensen, K. (2012). Sociomaterial approaches to conceptualising professional learning and practice. *Journal of Education and Work, 25*(1), 1–13. https://doi.org/10.1080/13639080.2012.644901

Flusser, V. (2011). *Into the universe of technical images. Electronic mediations* (Vol. 32). University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816670208.001.0001

Galloway, A. (2011). Are some things unrepresentable? *Theory, Culture & Society, 28*(7–8), 85–102. https://doi.org/10.1177/0263276411423038

Gant, S., & Reilly, P. (2018). Different expressions of the same mode: A recent dialogue between archaeological and contemporary drawing practices. *Journal of Visual Art Practice, 17*(1), 100–120. https://doi.org/10.1080/14702029.2017.1384974

Geimer, P. (2018). *Inadvertent images: A history of photographic apparitions*. University of Chicago Press.

Haraway, D. J. (2016). *Staying with the trouble. Making Kin in the Chthulucene*. Duke University Press.

Hayles, K. (2017). *Unthought: The power of the cognitive nonconscious*. The University of Chicago Press.

Historic England. (2018a). *3D laser scanning for heritage: Advice and guidance on the use of laser scanning in archaeology and architecture*. Historic England.

Historic England. (2018b). *Multi-light imaging highlight-reflectance transformation imaging*. Historic England.

Huggett, J. (2017). The apparatus of digital archaeology. *Internet Archaeology, 44*. https://doi.org/10.11141/ia.44.7

Huggett, J. (2020). Capturing the silences in digital archaeological knowledge. *Information, 11*(5), 278. https://doi.org/10.3390/info11050278

Huggett, J. (2023). Extending discourse analysis in archaeology: A multimodal approach. In C. Gonzalez-Perez, P. Martin-Rodilla, & M. Pereira-Fariña (Eds.), *Discourse and argumentation in archaeology: Conceptual and computational approaches*. Springer.

Hume, C. R., & Jones, G. W. (1959). Excavations on Nesscliff Hill. *Transactions of the Shropshire Archaeology Society, 56*(2), 129–132.

Huvila, I. (2017). Archaeology of no names? The social productivity of anonymity in the archaeological information process. *Ephemera, 17*(2), 351–376.

Huvila, I. (2019). Learning to work between information infrastructures. *Information Research, 24*(2), paper 819. Retrieved from http://InformationR.net/ir/24-2/paper819.html (Archived by WebCite® at http://www.webcitation.org/78mnTEFK7)

Huvila, I. (2022). Improving the usefulness of research data with better paradata. *Open Information Science, 6*(1), 28–48. https://doi.org/10.1515/opis-2022-0129

Huvila, I., & Sköld, O. (2021). Choreographies of making archaeological data. *Open Archaeology, 7*(1), 1602–1617. https://doi.org/10.1515/opar-2020-0212

Ingold, T. (2007). *Lines, a brief history*. Routledge.

Ingold, T. (2015). *The life of lines*. Routledge.

Jones, A. M., & Díaz-Guardamino, M. (2019). *Making a mark: Image and process in neolithic Britain and Ireland*. Oxbow Books. https://doi.org/10.2307/j.ctvjz80kw

Jones, J., & Smith, N. (2017). The strange case of Dame Mary May's tomb: The performative value of Reflectance Transformation Imaging and its use in deciphering the visual and biographical evidence of a late 17th-century portrait effigy. *Internet Archaeology, 44*. https://doi.org/10.11141/ia.44.9

Knight, M., & McFadyen, L. (2019). 'At any given moment': duration in archaeology and photography. In L. McFadyen & D. Hicks (Eds.), *Archaeology and photography: Time, objectivity and archive*. Bloomsbury Visual Arts.

Likavčan, L., & Heinicker, P. (2021). Planetary diagrams: Towards an autographic theory of climate emergency. In T. Dvořák & J. Parikka (Eds.), *Photography off the scale: Technologies and theories of the mass image* (pp. 211–230). EUP.

Lock, G., & Reilly, P.. (2019). Investigations resume at Nesscliffe Hillfort, Shropshire Archaeological and Historical Society Newsletter, 88.

Lock, G., & Reilly, P. (2020). Nesscliffe Hill Camp Excavations 2019, Shropshire Archaeological and Historical Society Newsletter, 89.

Lock, G., & Reilly, P. (2021). Nesscliffe Hill Camp: Fieldwork resuming in 2021, Shropshire Archaeological and Historical Society Newsletter, 91.

Martin-Rodilla, P., & Gonzalez-Perez, C. (2019). Metainformation scenarios in digital humanities: Characterization and conceptual modelling strategies. *Information Systems, 84*, 29–48.

May, J. (2019). *Signal, image, architecture (Everything is already an image)*. Columbia Books on Architecture and the City.

McFadyen, L., & Hicks, D. (2019). *Archaeology and photography: Time, objectivity and archive*. Bloomsbury Visual Arts.

Messemer, H. (2016). The beginnings of digital visualization of historical architecture in the academic field. In S. Hoppe & S. Breitling (Eds.), *Virtual palaces, part II. Lost palaces and their afterlife. Virtual reconstruction between science and the media* (pp. 21–54). Palatium e-Publications.

Miller, P., & Richards, J. (1995). The good, the bad, and the downright misleading: Archaeological adoption of computer visualisation. In J. Huggett & N. Ryan (Eds.), *CAA94. Computer applications and quantitative methods in archaeology* (pp. 19–22). Tempus Reparatum.

Min, J., Ahn, J., Ahn, S., Choi, H., & Ahn, S. (2020). Digital imaging methods for painting analysis: the application of RTI and 3D scanning to the study of brushstrokes and paintings. *Multimedia Tools and Applications, 79*, 25427–25439. https://doi.org/10.1007/s11042-020-09263-0

Mitchell, W. J. (1994). *Picture theory*. University of Chicago Press.

Mitchell, W. J. (2003). *M++: The cyborg self and the networked city*. MIT Press.

Mitchell, W. J. (2004). *What do pictures really want?* University of Chicago Press.

Morgan, C., & Wright, H. (2018). Pencils and pixels: Drawing and digital media in archaeological field recording. *Journal of Field Archaeology, 43*(2), 136–151. https://doi.org/10.1080/00934690.2018.1428488

Offenhuber, D. (2020). Data by proxy—Material traces as autographic visualizations. *IEEE Transactions on Visualization and Computer Graphics, 26*, 98–108.

Opgenhaffen, L. (2021). Visualizing archaeologists: A reflexive history of visualization practice in archaeology. *Open Archaeology, 7*, 353–377. https://doi.org/10.1515/opar-2020-0138

Opgenhaffen, L., Lami, M. R., & Mickleburgh, H. (2021). Art, creativity and automation. From charters to shared 3D visualization practices. *Open Archaeology, 7*(1), 1648–1659. https://doi.org/10.1515/opar-2020-0162

Ortolja-Baird, A., & Nyhan, J. (2021). Encoding the haunting of an object catalogue: on the potential of digital technologies to perpetuate or subvert the silencer and bias of the early modern archive. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqab065

Pelizza, A. (2021). Towards a sociomaterial approach to inter-organizational boundaries: How information systems elicit relevant knowledge in government outsourcing. *Journal of Information Technology, 36*(2), 94–108. https://doi.org/10.1177/0268396220934490

Pijpers, K. (2021). Worlding excavation practices. *Open Archaeology, 7*(1), 889–903. https://doi.org/10.1515/opar-2020-0177

Reilly, P. (1989). Data visualization in archaeology. *IBM Systems Journal, 28*(4), 569–579.

Reilly, P. (1991). Towards a virtual archaeology. In S. Rahtz & K. Lockyear (Eds.), *CAA90. Computer applications and quantitative methods in archaeology 1990* (BAR international series 565) (pp. 132–139). Tempus Reparatum.

Reilly, P. (1992). Three-dimensional modelling and primary archaeological data. In P. Reilly & S. Rahtz (Eds.), *Archaeology in the information age: A global perspective* (pp. 145–173). Routledge.

Reilly, P. (2015). Additive archaeology: An alternative framework for recontextualising archaeological entities. *Open Archaeology, 1*(1). https://doi.org/10.1515/opar-2015-0013

Reilly, P., Callery, S., Dawson, I., & Gant, S. (2021). Provenance illusions and elusive paradata: When archaeology and art/archaeological practice meets the phygital. *Open Archaeology, 7*(1), 454–481. https://doi.org/10.1515/opar-2020-0143

Reilly, P., & Dawson, I. (2021). Track and trace, and other collaborative art/archaeology bubbles in the phygital pandemic. *Open Archaeology, 7*(1), 291–313. https://doi.org/10.1515/opar-2020-0137

Reilly, P., & Lock, G. (2023, January/February). The Nessglyph: A horned iron age petroglyph? *British Archaeology*, 10.

Reilly, P., Lock, G., & Dawson, I. (2023). Preliminary observations on the 'Nessglyph', a petroglyph discovered at Nesscliffe Camp, Shropshire. *Transactions of the Shropshire Archaeological and History Society, 98*.

Reilly, P., Todd, S., & Walter, A. (2016). Rediscovering and modernising the digital Old Minster of Winchester. *Digital Applications in Archaeology and Cultural Heritage, 3*(2), 33–41. https://doi.org/10.1016/j.daach.2016.04.001

Rubenstein, D. (2020). The new paradigm. In D. Rubenstein (Ed.), *Fragmentation of the photographic image in the digital age* (pp. 1–7). Routledge.

Semper, G. (2004). *Style: style in the technical and tectonic arts, or, practical aesthetics* (Trans. H.F. Mallgrave & M. Robinson). Getty Research Institute.

Sköld, O., Börjesson, L., & Huvila, I. (2022) Interrogating paradata. In *Information research. Proceedings of the 11th international conference on conceptions of library and information science, Oslo Metropolitan University, May 29–June 1, 2022* (Vol. 27, special issue, paper colis2206). https://doi.org/10.47989/ircolis2206

Stewart, S. (1984). *On longing: narratives of the miniature, the gigantic, the souvenir, the collection*. Johns Hopkins University Press.

Taylor, J., & Dell'Unto, N. (2021). Skeuomorphism in digital archaeological practice: A barrier to progress, or a vital cog in the wheels of change? *Open Archaeology, 7*(1), 482–498. https://doi.org/10.1515/opar-2020-0145

Thomas, A. (2019). Duration and representation in archaeology and photography. In L. McFadyen & D. Hicks (Eds.), *Archaeology and photography: Time, objectivity and archive* (pp. 117–137). Bloomsbury Visual Arts.

Tringham, R. (2016). Chapter 16. Ruth Tringham with Michael Shanks and Christopher Witmore. In W. L. Rathje, M. Shanks, & C. Witmore (Eds.), *Archaeology in the making: Conversations through a discipline* (pp. 308–334). Routledge.

Wendrich, W. (2012). *Archaeology and apprenticeship: Body knowledge, identity, and communities of practice*. University of Arizona Press.

Witmore, C. (2007). Archaeology on the ground: the memory practices of David Webb Diggers Alternative Archive. *European Journal of Archaeology, 10*(1), 85–89.

Wittur, J. (2013). *Computer-generated 3D-visualisations in archaeology: Between added value and deception*. Archaeopress.

Zylinska, J. (2017). *Nonhuman photography*. MIT Press.

Zylinska, J. (2021). Undigital photography: Image making beyond computation and AI. In T. Dvořák & J. Parikka (Eds.), *Photography off the scale: Technologies and theories of the mass image* (pp. 231–252). Edinburgh University Press.

Ian Dawson is an artist, sculptor, lecturer, and researcher. He is director of the Critical Practices Research Group at Winchester School of Art and convener of the Materials Lab at the University of Southampton. With a background in sculpture Dawson's practice revolves around the intersection between 3D printing and new imaging. He often works as part of transdisciplinary collaborative teams to develop new ways to explore how 3D printing and imaging processes can be part of knowledge transfer processes. www.iandawsonstudio.com

Paul Reilly is an archaeologist and computer scientist, currently Senior Visiting Research Fellow in digital archaeology at the Department of Archaeology at the University of Southampton. He co-directs archaeological fieldwork at the Iron Age hillfort in Nesscliffe, Shropshire, UK. His theoretical research interests embrace creative digital archaeology and art/archaeology. His work explores some of the implications of the growing intersection of physical and digital (or phygital) practices for art, archaeology, and cultural heritage.

Mapping Accessions to Repositories Data: A Case Study in Paradata

Kevin Matthew Jones and Jenny Bunn

Abstract

Accession is the term used within archive services to refer to both the process of taking 'intellectual and physical custody of materials' and to those materials themselves—'the materials physically and officially transferred to a repository as a unit at a single time' (Society of American Archivists. Dictionary of Archives Terminology, s.v. 'accession'. Retrieved December 15, 2022, from https://dictionary.archivists.org/entry/accession.html, n.d.). This chapter describes an exercise to repurpose data collected about accessions to repositories over the period 2007–2020. The steps involved in cleaning and preparing the data are described and a contextual narrative to them is provided. Reflections are offered on the idea of paradata as it was explored and put into practice during the project.

1 Introduction

This chapter centres on a research project based on the re-use of data detailing the material accessioned into archival institutions in the UK during the period 2007–2020. This project required the assemblage of both these data and an understanding of that data in order that it might become of use for a particular purpose different to that for which it was intended when it was originally collected. In this way, data

K. M. Jones
The University of Leeds, Leeds, UK
e-mail: K.M.Jones@leeds.ac.uk

J. Bunn (✉)
The National Archives, London, UK
e-mail: Jenny.Bunn@nationalarchives.gov.uk

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_7

became dataset as data were re-set and re-framed in order to serve this new use. The processing whereby the data were set will be described, as will the information which was gathered and/or recorded as part of this process of re-setting. During this process, the concept of paradata was encountered and the work being undertaken provided an opportunity to explore it further. Questions arose of whether, in the context of this work, such a concept was relevant and, if so, what, from that work, might be considered paradata. Reflections on these questions will be offered at the end of the chapter, in order to encourage readers to shape their own reflections on paradata as it applies to their own work and context.

The National Archives (TNA) holds a wealth of accumulated data concerning archival materials held across the UK and beyond. The existence of this data is the result of concerted collective effort by a number of bodies and projects over many years. These bodies have included staff at archive services operating at both local and national levels and projects including the National Register of Archives (NRA) (which was started by the Historical Manuscripts Commission shortly after the Second World War) and the Access to Archives project (which ran at the turn of the twenty-first century). The data used by the project described in this chapter are the result of the National Accessions to Archives Survey, which TNA runs annually. Accession is the term used within archive services to refer to both the process of taking 'intellectual and physical custody of materials' and those materials themselves—'the materials physically and officially transferred to a repository [or archive service] as a unit at a single time' being described as an accession (Society of American Archivists, n.d.). The National Accessions to Archives Survey requests archive services around the UK to provide brief details of all the accessions they have received within a calendar year.

To date, the main use and principal reason for the collection of this data has been to enable researchers to discover and locate sources of interest around the UK. However, a growing realisation that this data also held the potential to produce insights into both historical and contemporary patterns of collecting across the UK led to a desire to try to realise this potential. Focusing as much on investigating the possibilities and mechanisms for generating insights as on the insights themselves, a 12-month Research Fellowship was established in 2021 at The National Archives to re-assemble a subset of the data gathered during the annual survey exercise (hereafter referred to as the accessions to repositories data) and then to explore its potential for surfacing insights around patterns of archival collecting. The Fellowship concluded at the end of 2022, and further details of the work carried out during it can be found in the following section.

2 Starting to Set the Data

At the start of their work in 2021, the Fellow was presented with the data that were both readily available and thought to be of potential use given the aims of their work. These data consisted of details of annual accessions and existed in the main in two distinct forms. The first form existed as a set of csv data frames that each contained information on accessions for the years 2016–2020. These data frames had merged the annual accessions information supplied by repositories across the UK from these years. The framing of this data had been carried out according to the Tidy Data principles outlined by Hadley Wickham and others, in that each column in the data frame was a variable (2019). The content of these columns was defined in terms that included ARCHON Number (a unique identifier for the individual repository or archive service into which the material had been accessioned), Name of Repository, Record Creator (who had created the material being accessioned), Brief Description of the Record (what that material consisted of), Size of the Record (how much material there was), and the Dates Covered by the Record (the time period represented by/within the material).[1]

Data from the years before 2016 were in not such an easily accessible form but rather in one that prevented analysis and the identification of patterns in collecting practices. The data detailing the accessions between the years 2007 and 2015 had been stored on TNA systems as a series of files that were in the original format returned by individual repositories participating in the annual accessions to repositories survey. These mainly took the form of an Excel file sent out by representatives of TNA to participating institutions. This form also collected data described in the previous paragraph, but since it had changed format and layout since 2007, a simple merging of the forms was not possible. Furthermore, many repositories had submitted their return to the survey in many other formats, including Word documents, .pdfs, and in some cases as image files of printouts. These were themselves attached to emails stored as MS Office Outlook (.msg) files in folders within the corporate Sharepoint system. The framing of data between 2007 and 2015 was therefore considerably less helpful with respect to its state of readiness for subsequent use or analysis.

The data from between 2007 and 2015 therefore needed lengthy processing in order to get them into the same state as the data for the years 2016–2020. The first lengthy task was to extract the attachments from the stored emails. Once this procedure had been completed, the next task was to bring all the data contained in these many different attachments together into a single data frame for each individual year. In both cases this required a largely manual, non-computational effort, which was simultaneously extremely time-consuming and repetitive, yet necessary to lessen the risk of accidental data loss or misinterpretation.

Issues encountered during this process included the discovery that differing interpretations of what should be entered into the different columns had led to problems with a small yet significant number of the returns. For example, in many returns, information had been included in the wrong place, such as details of record donors included within record creator fields, or the covering dates becoming confused with the date of record creation. A number of strategies were developed to identify, and where possible rectify, these problems with the data. In cases of

[1] The spreadsheet template also included a number of other fields including the url of the record if available, and whether the record was an addition to an existing collection.

the data being inserted into the wrong column, it was possible to use data wrangling programmes such as Google Refine and the Python library PANDAS to help identify and move the data. In another significant number of returns, the year to which they referred was not self-evident, and a judgement call needed to be made. This was because many institutions would submit a number of returns at once. For instance, one repository had not sent in returns between 2007 and 2010, but had sent through all the returns in 2011 for that year, as well as the preceding four. Since the goals of the Fellowship were to use the survey data to understand trends and patterns in accessions, the decision was taken in this and similar cases to separate the data by year of accession and not by the date it was received by TNA.

Another aspect of processing that shaped the final dataset concerned the standardisation of the data within it. Substantive work was carried out to standardise many common values, such as the names of prominent individuals, dates, and locations appearing in the individual accession descriptions. Hardly any of the fields in the standard returns template were authority controlled and the different styles of those completing them led to much variety. For example dates were written in a number of different formats—'Jun 1915', 'June 1915', '01/06/1915'.

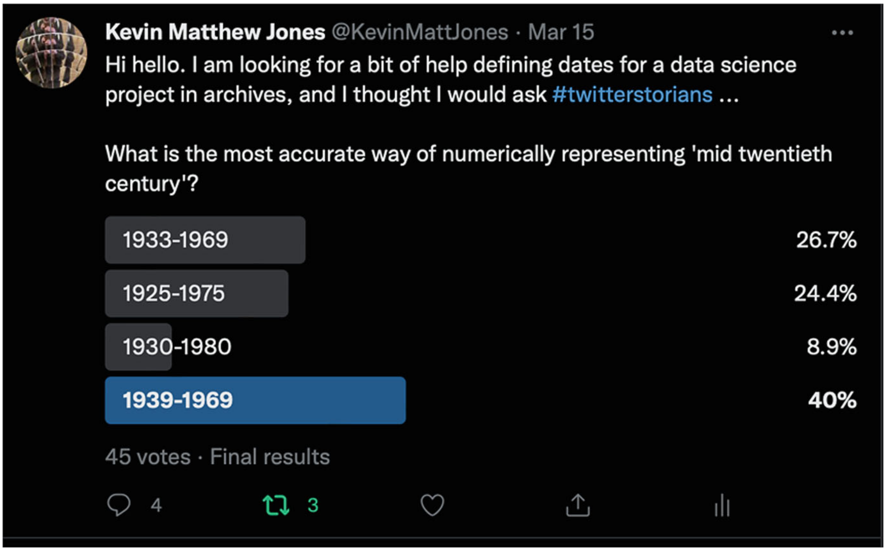
To make the data machine readable, the Fellow decided to replace natural language phrases such as 'mid-twentieth century' and 'Victorian England', with numerical representations in the format yyyy–yyyy.[2] The nearest thing to a UK archival standard on how to do this was contained within the National Council on Archives Rules for the Construction of Personal, Place and Corporate Names, which offers the following model (National Council on Archives, 1997):

Early nineteenth century	1800–1840
Mid-nineteenth century	1830–1870
Late nineteenth century	1860–1899

However, before applying this model, the Fellow wanted first to ascertain whether it was one that would be shared by others. He therefore sent out a tweet to ask the question, 'what is the most accurate way of numerically representing 'mid-twentieth century'?' (Fig. 1). The tweet received 45 responses, and whilst not a scientific measure, these responses demonstrate a decided difference of opinion on how 'mid-century' should be numerically represented. It was initially thought that xx25–xx75 would come as the most favoured choice because it rested on an intuitive division of a century into quartiles,[3] but as it turned out, it was the second least

[2] The Fellow decided to remove information on days and months because given the wealth of information, it would not be possible to represent these effectively within sector wide representations of accessions data that emerge from the project.

[3] The quartiles being early = xx00 – xx24, mid or middle being xx25 – xx74, late being xx75 – xx99.



Kevin Matthew Jones @KevinMattJones · Mar 15

Hi hello. I am looking for a bit of help defining dates for a data science project in archives, and I thought I would ask #twitterstorians ...

What is the most accurate way of numerically representing 'mid twentieth century'?

1933-1969	26.7%
1925-1975	24.4%
1930-1980	8.9%
1939-1969	40%

45 votes · Final results

Fig. 1 Results of a poll carried out on Twitter as to the most accurate way of numerically representing mid-twentieth century

popular choice amongst researchers on Twitter, and some complained that 1933–1968 was not included as an option.[4]

Bowker and Star explore the problematic nature of standardisation and formalisation within the context of formal medical nosology, namely the World Health Organization's International Classification of Diseases (2000). They provide an ideological reading of the medical nosology, claiming that the act of codification strips social and economic factors from the conditions it lists. A concern during this fellowship was that the standardisation required by the processing/purpose of the work would unwittingly remove meanings, conscious or unconscious, from the data in the returns, reducing or stripping entirely the agency of those who completed them. The need to make judgements on how to standardise data like dates had to be balanced with a desire not to lose meanings contained in diversity of expression. A number of compromises between diversity of expression and the need to make the dataset machine readable needed to be made during the work to clean the data. To improve transparency, a document that outlined and discussesd the decisions taken to standardise and therefore shape the data was produced in the form of a lengthy report. For instance, due to the condition of some of the returns and the time constraints upon the Fellowship, it was not possible to avoid all data loss during the processing described above. The amount of data that were lost is estimated at around

[4] Some viewed Hitler coming to power and May 1968 as the beginning and end of the mid-century. Others believe the 'mid-century' to have begun with the Wall Street crash in 1929 and to have ended with the oil crisis in 1972.

1% for each year. This was judged to be statistically insignificant for the exercise, but it is nonetheless a loss. The whole process of assembling and cleaning a dataset that consolidated and standardised the available accessions to repositories data from 2007 to 2020 took approximately 6 months.

3 Understanding the Data

Whilst the data were being set, and decisions were being taken about how it might best be standardised, attempts were also being made to understand better what it represented and where it had come from. As a result the following contextual narrative was constructed.

The annual accessions to repositories survey traces its origins back to 1923 and to a practice initiated at that time of the regular publication within the *Bulletin of the Institute of Historical Research* of listings of the movements of historical manuscripts. These listings were constructed from two main sources, firstly catalogues detailing the sale of manuscripts and secondly annual reports from local and national repositories, which provided details of manuscripts they had received during the year. By the early 1950s however the *Bulletin* was finding it difficult to allocate sufficient space to these ever-growing listings and overlap was also starting to be seen with the work of the National Register of Archives which had been established under the aegis of the Historical Manuscripts Commission (HMC) in the immediate post-Second World War period. In 1955, the first HMC List of Accessions to Repositories was published, providing a summary of those details of accessions that had been received from 86 repositories across the UK (Historical Manuscripts Commission, 1955, p.1).

The process of compiling this annual listing was framed very much as an editorial one. Prefatory and editorial notes often preceded the published listings, and these notes provide insight into some of the decisions being made as part of that process. For example, in the listing for 1955 it was noted that: ‘it has again been necessary to curtail some of the reports considerably, and much detail has had to be omitted’ (Historical Manuscripts Commission, 1956, p. 1). Then again, the listing for 1963 noted;

> In order to control the increasing size of this *List*, certain entries have been given in less detail than hitherto, viz,:
>
> a) *Deeds* are not described in detail when they relate to the area where the repository is situated; details are however given when they relate to other areas.
> b) *Parish* and *school records* are not described in detail unless the number of such deposits shown is few or the records are of particular interest (Historical Manuscripts Commission, 1965, p.iii).

The editors of the list seem to have faced a constant battle to keep the publication to a ‘reasonable size’ as the number of repositories approached for returns steadily increased to over 200 by the mid-1990s (Sargent, 1995).

As the size of the volume grew, changes were also made in terms of arrangement, indexing, and layout to make the listings easier to consult. An alphabetical ordering by repository name was often favoured, although indexing repository to geographic county was also considered important, leading to various editorial notes on changing county boundaries over the years. A major typographical change in the listing for 1972 seems to have reflected a move towards an even more selective approach towards compilation. It being noted that:

> Efforts are also being made to simplify the contents as far as possible by aiming to provide outline descriptions only of the more important accessions to each repository in place of what had become a welter of indiscriminate and undigested detail (The Royal Commission on Historical Manuscripts, 1974, p.v).

The annual listings were not the only way in which accessions to repositories information was published. Even more selective lists or digests—of those accessions relevant to researchers of particular subjects—were also commonly supplied for onward dissemination, often through publication in the bulletins or journals of research societies and communities. It is reported that digests of accessions for 1992 were supplied to around 30 such organisations.

The year 1992 also marked the discontinuation of the hard copy publication of the annual listings (Sargent, 1995). The Historical Manuscripts Commission—whose staff managed the annual accessions exercise—had eagerly embraced the coming of the computer and was quick to create its own presence at the birth of the Internet. Online publication of the listings of accessions (both the ‘full’ and subject-specific digests) became the norm from then on, and examples of this are preserved in Web Archives (The Royal Commission on Historical Manuscripts, 1995). Copies of the subject-specific digests did continue to be submitted to (and subsequently published in hard copy) in academic journals after that date, but eventually this practice also came to an end.

The information gathered in the regular accessions to repositories exercise never fed solely into the editorial process that led to the annual lists and digests of the same; rather it had long also fed into the wider work of the National Register of Archives. This body sought to act as a central gateway to and (even before the term gained meaning) database of UK archives. As well as the accessions to repositories returns, it also gathered in (and in some cases created and published) surveys and more detailed finding aids, as well as creating and maintaining master indexes to the same. New accessions engendered new index entries, which were in due course enhanced through connection to more detailed finding aids as the information held about UK archives by the National Register of Archives grew over time.

Gathering, holding, manipulating, and linking all this information in the pre-digital era involved a lot of paper and human labour in copying information across from returns to index cards and other such technologies. The potential for computerisation started to be explored in the mid-1980s and a bespoke system based on a Prime minicomputer and programs written in Ampersand Pace came into operation in 1987 (Sargent, 1995). By the mid-1990s, this system was migrated into the newer Windows/PC environment and contained eight databases: indexes to the people, organisations, families, and estates about which UK archives were held and some details of the location of manorial documents, of archive repositories, and of listings and finding aids received from them (Sargent, 1995). Information from accessions to repositories returns (mostly still received in a paper form) was selected to enhance (and manually input directly into) the indexes, but it also had to be selected, added (and manually input again) into a separate system/database to create the annual subject-specific and all repository listings.

The next big system change occurred around 2002/3 at the same time as the bringing together of HMC with the Public Record Office and Her Majesty's Stationery Office to create The National Archives. This system, HMC Admin, was more interconnected, such that the need for double entry of information from the accessions to repositories returns was reduced. An explanation of the process through which information was added into the HMC Admin system from around 2012 runs as follows:

> **What happens to my return after it is submitted and acknowledged?**
> Your return is logged and saved into our electronic file management system. The Accessioner for your region (we have divided the UK and Ireland into 14 separate regions) reviews your returns and makes selections for inclusion into the NRA and Accessions. They then input the information into the NRA and tick a box in the NRA entry, which adds it to Accessions. On completion the Accessions Editor proof reads each Accessions entry and assigns it to a thematic digest if applicable. Once all of the entries have been checked and assigned to a relevant digest the complete survey is published online at http://www.nationalarchives.gov.uk/accessions/. (The National Archives, c.2012).

It is undeniable that there are a lot of choices and assumptions underlying the accessions to repositories data. The process into which it fed has prioritised its digestion over its analysis. As we saw above, there was a constant drive to surface only the most significant and interesting accessions—to facilitate use by providing the most pertinent selection from the 'welter of indiscriminate and undigested detail'. This selection was made not only by those centrally processing the returns, but also by those submitting them, as the following extract from the editorial note to one of the early accessions listings makes clear:

> Many repositories have assisted greatly by submitting their lists in a concise form suitable for publication with little alteration; it would be appreciated if others could do the same, as the selection of the most significant facts or items is far more easily done by the actual custodians of documents than by an editor at a distance (Historical Manuscripts Commission, 1956, p.1).

These multiple acts of selection are not well-documented, at least not at anything below the level of general principles as set out in the editorial notes and quoted earlier.

This absence is particularly felt in the case of data relating to accessions to repositories, because accessioning is itself a form of selection, the decision to accession and hence preserve certain records or not. It is here that the work of archivists can have its most powerful impact. In this decision lies the potential for continuing to perpetuate structural inequalities whereby the margins are rendered all but invisible to history. This potential has long been recognised. For example, writing in the 1970s, Felix Hull wrote that:

> Quite bluntly so long as we are not involved in selection we can happily be all things to all archives, but once we assume the sword and scales of justice—what then? [. . .] For some archivists, I am sure, this dilemma raises problems of action and of morality which they feel ill equipped to handle (Hull, 1979).

The archival profession is still wrestling with those problems of action and of morality that Hull described today, and it is for this reason that the decision was made to focus on the accessions to repositories data in this project. In theory, it held the potential to reflect back an evidenced picture of what had been collected in the past that could at least provide some more robust information for those continuing to wrestle with those problems. In practice however, its selective nature—the fact that it was a selection—raised concerns about the extent to which it could be relied on as such a picture. That selection may have served the purposes of the original use—to produce a digest or summary of recent accessions of significance for the benefit of researchers—but it did not serve the purposes of the data’s re-use—to undertake analysis of patterns of collecting in UK archival repositories over time.

4 Reflecting on Paradata

Unlike the term metadata, paradata does not yet have currency within the archives field and neither of the researchers involved in the above project were previously experienced in thinking in its terms. When prompted to do so, by an invitation to contribute to this volume, they undertook an initial investigation into its conceptualisation elsewhere, identifying (as others in this volume and elsewhere have already done) a number of origin stories and extant definitions.

The origin story and conceptualisation that resonated with them the most traced back to the coining of the term ‘paradata’ by Drew Baker during the course of the Making Space project carried out at King’s Visualisation Lab to investigate ‘a methodology for tracking and documenting the cognitive process in 3-dimensional visualisation-based research’ (King’s Visualisation Lab, c.2005). In this case, the concerns from which the need to conceptualise paradata arose were open questions around the credibility and validity of 3-D visualisations within the archaeological and wider arts and humanities research communities. Such concerns led not only to the coining of the term paradata, but also to the London Charter for the Computer Based Visualisation of Cultural Heritage, which ‘defines principles for the use of computer-based visualisation methods in relation to intellectual integrity, reliability, documentation, sustainability and access’ (2009a). It too provides a definition for paradata, the glossary stating that it is:

> Information about human processes of understanding and interpretation of data objects. Examples of paradata include descriptions stored within a structured dataset of how evidence was used to interpret an artefact, or a comment on methodological premises within a research publication (London Charter, 2009b).

It is also via this path that a connection has been initiated with the archival field. This connection taking the form of Heidi Jacobs of the University of Windsor who references the London Charter and interprets paradata as ‘a way to reveal the “fingerprints” of those who created the heritage object and the choices and assumptions that led to its creation’ (2020).

This conceptualisation of paradata resonated with the work being conducted on the accessions to repositories data because of this focus on ideas of fingerprints, choices, and assumptions. As has been discussed above, not only were there lots of current choices being made in the reassembly and standardisation of these data, researching the history of them also led to many more choices and assumptions. Those of the archivists who not only decided what material they would or would not accession in any given year, but also whether, or not to complete the survey at all, and to what level of detail. Further choices were made by the ‘editors’ as they decided what they felt to be important enough for inclusion in the published digests. Practicalities such as the technology and time they had available to them played a role in such decisions, but clearly assumptions were also being made, assumptions which were not always consciously acknowledged at the time.

In many ways it was these assumptions, particularly those about what was or was not considered important, which the Fellowship had set out to uncover, looking through and working backwards from the data on what had been collected to try to draw inferences, or at the very least prompt reflection, about what had been considered important—the assumptions on which archivists of the past appeared to have been working in their selection of material for inclusion in their collections. Perhaps it was because of this—our purpose for the data in question—that the concept of paradata became so resonant? Perhaps paradata only arises as a concern when you seek to look through the data to infer something else from it, when you seek to understand if the drawing of such an inference is justified, evidenced, or even possible? Given then that this concern had arisen in relation to the project, how did we seek to address it? What did we find or use or produce ourselves that could, as a consequence of meeting the concern with paradata be considered to have acted as such, to be paradata?

Meeting the concern with paradata required information; it required the gathering and taking into account of information about (a) the processes and choices by which the data had originally been collected, (b) the way in which our own processing of it was reshaping it and leading to some data loss, and also (c) the extent to which the dataset could be seen to be complete and/or representative of what we were using it to stand in for—the pattern of collecting undertaken by archive repositories in the UK over the period 2007–2020. This information was sourced in many ways, from previously published articles and book chapters, internal TNA reports and documents, and the knowledge of those who had been involved in the work of originally collecting and processing the data, which was sometimes gathered through conversation with them and sometimes through the editorial notes they had left behind. Then again, it was also gathered by consideration of the gaps between what was represented in the dataset (e.g. in terms of the number of repositories who had made returns in any given year) and what was not (e.g. in terms of how that number differed from what was known of the total number of repositories in operation). It was on the basis of this sort of information that we worked with the dataset and in our placing of it alongside that dataset, outside it and yet fundamental to its interpretation and use, it can, from our perspective, all be considered paradata.

In many ways then, this view of paradata paints it as something similar perhaps to another concept, that of a knowledge base as defined by the Reference Model for an Open Archival System as 'a set of information, incorporated by a person or system, that allows that person or system to understand received information' (International Standards Organisation, 2012). As part of this project, we put much effort into assembling our knowledge base, the information that allowed us to use and interpret the data with which we had been presented. To be sure we incorporated this knowledge, but we also took pains, in our final reporting of the project to dis-incorporate or rather to disembody it, to set it out and alongside in sufficient detail that anyone coming across that data in the future would at the very least be able to work with them on the same base or basis that we had.

5 Conclusion

This chapter has focused on a project which aimed to re-use data (dating from 2007 to 2020) collected as part of the annual accessions to repositories survey in order to create a picture of collecting patterns across UK archives. The original process of collecting the data was outlined as was the state of the data when it was first encountered. The work involved in preparing this data for its new use was described, highlighting the importance of standardisation, and the history reconstructed around the data was also set out. Leading on from this, reflections were offered on how paradata came to be conceptualised during the course of the project. Looking into earlier conceptualisations, the idea of paradata was seen to resonate with the project in its concern with looking through that which was being considered as 'the dataset' to something beyond—the drawing of inference or conclusion from it. In order to achieve this goal, it was necessary within the project to seek out a range of additional information beyond the original dataset. This information as well as knowledge of the process being followed to reshape the dataset came to act as paradata in that it addressed a concern with the basis or base on which the data could be used and interpreted. In the final reporting of the project then, a selection was made that prioritised that information it was felt necessary to pass on in order that the dataset could be used on the same basis in the future. This reflection and conceptualisation has been offered to encourage readers to shape their own reflections on paradata as it applies to their own work and context. It is also hoped that readers will be encouraged to consider what information they should pass on alongside any dataset they define or set in order that others can either understand, interpret, and use it on the same basis, or be aware of how they are not doing so.

References

Bowker, P., & Leigh-Starr, S. (2000). *Sorting things out: Classification and its consequences*. MIT Press.

Historical Manuscripts Commission. (1955). *Bulletin of the National Register of Archives No. 6*. Her Majesty's Stationery Office.

Historical Manuscripts Commission. (1956). *Bulletin of the National Register of Archives No. 8: List of accessions to repositories*. Her Majesty's Stationery Office.

Historical Manuscripts Commission. (1965). *List of accessions to repositories in 1963*. Her Majesty's Stationery Office.

Hull, F. (1979). The archivist and society. *Journal of the Society of Archivists, 6*(3), 125–130.

International Standards Organisation. (2012). *Space data and information transfer systems—Open archival information system (OAIS)—Reference model* (ISO Standard No. ISO 14721:2012). https://www.iso.org/standard/57284.html

Jacobs, H. (2020). Invisible in plain view: Libraries, archives, digitization, memory, and the 1934 Chatham Coloured All-Stars. In M. Kandiuk (Ed.), *Archives and special collections as sites of contestation* (pp. 223–248). Litwin Books & Library Juice Press.

King's Visualisation Lab. (c.2005). *Making space*. King's College London. Retrieved September 27, 2022, from https://www.kvl.cch.kcl.ac.uk/making_space.html

London Charter for the Computer Based Visualisation of Cultural Heritage. (2009a). *Preamble*. Retrieved September 16, 2022, from http://www.londoncharter.org/preamble.html

London Charter for the Computer Based Visualisation of Cultural Heritage. (2009b). *Glossary*. Retrieved September 16, 2022, from http://www.londoncharter.org/glossary.html

National Council on Archives. (1997). *Rules for the construction of personal place and corporate names*. Retrieved September 27, 2022, from https://archiveshub.jisc.ac.uk/ncarules/

Sargent, D. (1995). The National Register of Archives. In D. Sargent (Ed.), *The National Register of Archives: An international perspective essays in celebration of the fiftieth anniversary of the NRA* (pp. 1–35). University of London Institute of Historical Research.

Society of American Archivists. (n.d.). Dictionary of Archives Terminology, s.v. "accession". Retrieved December 15, 2022, from https://dictionary.archivists.org/entry/accession.html

The National Archives. (c. 2012). *Accessions to repositories FAQs*. Unpublished.

The Royal Commission on Historical Manuscripts. (1974). *Accessions to repositories and reports added to the National Register of Archives 1972*. Her Majesty's Stationery Office.

The Royal Commission on Historical Manuscripts. (1995). *Accessions to repositories in 1994*. Retrieved September 27, 2022, from https://web.archive.org/web/19980704014639/http:/www.hmc.gov.uk/accessions/1994/94digests/intro.htm

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the Tidyverse. *The Journal of Open Source Software, 4(43)*, 1686. https://doi.org/10.21105/joss.01686

Kevin Matthew Jones Ph.D., is lecturer in the History and Philosophy of Science at the University of Leeds. He teaches science communication and topics in the history of modern science. He has held research fellowships at The National Archives, the University of Birmingham, and the Leeds Art Humanities Research Institute. He is currently working to convert his Ph.D. thesis on the history of public health data and mental health statistics for publication as a part of Palgrave MacMillan's Mental Health in Historical Perspective series.

Jenny Bunn Ph.D., is Head of Cataloguing, Taxonomy and Data at The National Archives (UK). She has thirty years' experience as an archival practitioner, educator, and researcher. Her research interests lie at the intersection of archives and technology, both from a historical perspective and looking forward to the implications of artificial intelligence for archival practice.

Paradata for Digitization Processes and Digital Scholarly Editions

Wout Dillen

Abstract

As libraries and archives are increasingly digitizing their collections, their resulting digital reproductions are now also reused in various research outputs. Because their patrons typically come from diverse backgrounds, however, many of them lack the necessary experience with the intricacies of the digitization process to judge how this process may have affected the quality of the reproductions they intend to (re)use. Without easily comprehensible paradata (i.e., data that indicates how they were made), patrons have no choice but to take these digital objects at face value—which is a problematic research practice. To illustrate some of the ways in which the digitization process may affect the reproduction, this chapter discusses a case study where a researcher commissioned the digitization of a collection of manuscripts held by various memory institutions across Sweden. By zooming in on how quality standards are negotiated between researchers and library staff in a specific digitization project, and the problems they needed to resolve along the way, this chapter examines which types of paradata could be useful to contextualize digitization processes and gives a concrete suggestion how the reusers of those digital reproductions could in turn provide essential paradata to contextualize their own research outputs.

W. Dillen (✉)
Högskolan i Borås, Borås, Sweden
e-mail: wout.dillen@hb.se

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_8

1 Introduction

In the last few decades, the “digital turn” has drastically changed the way in which at least the Western world interacts with the cultural documents of its past. As more and more memory institutions (such as libraries and archives) have begun to digitize the cultural heritage documents in their collections, researchers (as well as the general public) interact less with the original artifacts directly, and more indirectly with their readily accessible digital representations. Since libraries and archives can essentially function as meeting places for researchers from any disciplinary background, and the documents in their collections are invariably consulted for a myriad of different purposes, only a fraction of their patrons will have a nuanced understanding of how the digitization of their source materials were performed, and how this largely invisible process may impact their research. Building on Isto Huvila’s definition of paradata in his recent publication on “Improving the Usefulness of Research Data with Better Paradata,” such “contextual knowledge about how data [in this case: the digital reproduction] was created and how it has been curated and used” is called paradata (Huvila, 2022, p. 28). Precisely which types of data and workflows are useful in this context will be discussed in more detail below. At this point, however, we would do well to remember that the availability of paradata (in general) is especially “critical when (re)users come from diverse disciplinary backgrounds and lack a shared tacit understanding of the priorities and usual practices of obtaining and processing data” (Huvila, 2022, p. 30; paraphrasing Doran et al., 2019). This is often the case when memory institutions share their digitized documents with their patrons. In the absence of paradata, those patrons have no choice but to take the digital representations at face value, unless they meticulously check each digital image against its original—thereby largely defeating the purpose of the reproduction. The lesson here is not so much that we should distrust the library or archive and the materials they make available, but rather that we need more data to build this trust on, if we want to reuse those materials in a research setting. As I will argue in this chapter, this issue is especially relevant in the fields of textual scholarship and its more practice oriented sister discipline “(digital) scholarly editing,” where researchers develop what are supposed to be critically informed authoritative editions of cultural heritage documents—although they rarely justify the relation between the original physical documents and their digital reproductions when doing so.

By having a closer look at a workflow for digitizing cultural heritage documents across a selection of memory institutions from a researcher’s perspective, this chapter will consider questions such as: what kind of problems can occur while photographing cultural heritage documents, and how may these problems influence the researcher’s results? What kind of paradata would be useful when we try to contextualize these types of digitization processes? And what kind of paradata would we need to provide when we reuse those digitized cultural heritage documents to develop our own digital scholarly editions? As such, the chapter makes a key contribution toward facilitating paradata provision in this area in the future. Still, while the chapter is based on the analysis of a selection of specific case studies, their results and conclusions are potentially useful in the broader area of information and knowledge management research and practice, especially in relation to those areas of scholarly and professional work where digital reproduction mechanisms and the application of critical and forensic perspectives to the organization of knowledge and information are crucial to the task at hand.

2 Textual Scholarship and the Digital Reproduction

Before doing so, however, I should first introduce the two academic disciplines that will frame our discussion of paradata in this chapter: textual scholarship and scholarly editing. Both these disciplines share a crucial focus on textual documents of cultural, historical, or political significance that are being preserved, cataloged, and made accessible in memory institutions across the world. Studying these documents (i.e., analyzing the differences and similarities between versions, what this information implies about how they have been transmitted over time, and how this knowledge may inform our understanding of those documents) falls within the purview of the field of textual scholarship. Applying this knowledge and visualizing it in the form of an authoritative edition that contextualizes such documents falls within the purview of the field of scholarly editing.[1] Both disciplines are sometimes looked down on as "auxiliary sciences" in academia—useful tools that nevertheless mainly serve to lay the groundwork for the "real" research. Formulated more positively, it has been argued that the theory and practice of scholarly editing is the bedrock of many "disciplines in the humanities, such as literary studies, philology, history, philosophy, library and information science, and bibliography" (Boot et al., 2017b, p. 15).

Whichever way you look at it, these disciplines provide us with some much needed context to the objects of our research—objects that are increasingly being digitized.[2] And these digital surrogates are also increasingly used in different stages of textual scholarship and scholarly editing workflows (Dillen, 2017b). It frequently occurs at the research stage, for example, when the scholar uses digital facsimiles to transcribe the texts of their source materials, interprets them, collates and compares different versions of the texts, etc. Ideally, of course, the scholar will not rely solely on these digital surrogates for their research, but rather complement their workflow with visits to the archives to inspect the original source materials, familiarize themselves with the materiality of the documents (which may provide

[1] For examples of digital scholarly editions, see Patrick Sahle and Greta Franzini's respective catalogs: Sahle (2008) and Franzini et al. (2016). For novices to the field, see my introductory video lecture on "What is a Digital Scholarly Edition?" (Dillen, 2022).

[2] For more information on the digital turn in scholarly editing, see: Sahle (2013); Apollon et al. (2014); Pierazzo (2015); Driscoll and Pierazzo (2016); Boot et al. (2017a).

new insights), and double-check their findings.[3] But here too, the fact remains that the bulk of the research time is spent handling digitized objects rather than their physical originals—which implies that a lot of trust is placed in their quality and representativeness.

This trust becomes even more critical in the reception phase. Once the digital facsimiles are published as a digital scholarly edition—an environment where they apparently receive an implicit seal of approval from subject experts in the field—they effectively become the default point of entry for the users of the digital scholarly edition, who are highly unlikely to compare each digital facsimile to its archived original. Which also means that it is exactly this digital facsimile that will be used as a point of reference to evaluate the accuracy of the editor's transcriptions and the validity of the claims and arguments they are proposing about their source materials.[4] As such, this system enables researchers to push the responsibility for the quality of the digitized materials forward—from the critical user to the editor, and from the editor to the archive.

The fact that the users of digital editions are encouraged to take the edition's facsimiles at face value in this way is ironic, to say the least, because textual scholarship, as a discipline, is centered around the assumption that when a text is transported from one carrier to the next, variance is likely to occur. And that it is exactly through the study of this variance that we can reconstruct the history of the text, learn how it has been transmitted over time, and how its contents and reception may have changed in the process. This is doubly true when a text is transported from one medium onto another, which will invariably alter our interaction with, and hence to some extent our understanding of, the text in some minor or major way. This was the case when our practices shifted from manuscript to print, and again when those handwritten or printed documents are digitally photographed. The fact that such digital photographs are never truly "unedited" or objective was already convincingly made by textual scholar Hans Zeller's in the 1970s (Zeller, 1971). More recently, the reproductive potential of digital surrogates was questioned by Lars Björk, senior conservator at the National Library of Sweden (Björk, 2015). As Björk argues, when we neglect the complexities of transmission, and uncritically take for granted that the reproduction is "capable of replacing direct access to collections and documents, we will risk basing our knowledge and assumptions on sources that have a restricted capacity to convey the information potential of the document" (Björk, 2015, p. 5). Yet this is arguably exactly what is happening in the fields of textual scholarship and digital scholarly editing.

The problem here is that while the digitized images of source documents have become an increasingly significant component of the digital scholarly edition over the last decades, scholarly editors are largely unaware of the intricacies of the

[3] Of course, this ideal is not always attainable—especially for scholars who lack institutional embedding or have limited funds.

[4] For discussions on how digital scholarly editions are used to construct an argument about their source materials, see: Andrews and van Zundert (2018); Bleeker and Kelly (2018); Dillen (2018).

digitization processes that are being practiced at memory institutions across the world, and rarely (if ever) document them in their editions. This systematic lack of paradata is problematic because it does not account for the fact that the quality of the results of these digitization processes is largely determined by the availability of resources and expertise and the successful negotiation of terms with third parties that memory institutions employ to outsource some of their digitization needs. Indeed, the results of my studies across European National Libraries as part of the DiXiT project[5] confirmed our hypothesis that the digitization workflows and standards that are practiced in individual memory institutions were flexible and constructed through a process of negotiation that involved various parties that each brought their own stakes and perspectives to the table (Dillen, 2017a). It is my hypothesis that documenting our interaction with these standards, practices, workflows, and decisions in the form of paradata can help fill an existential lacuna in the contextualization of our research documents in digital scholarly editions and thereby also play an essential role in convincing the end user of the validity of our work.

3 Case Study: manuscripta.se

To back this hypothesis up, I will focus on the findings of my main case study in the DiXiT project: the National Library of Sweden (hereafter: KB-SE). Specifically, I will reflect on an interview I conducted with one of its resident researchers: Patrik Granholm, creator of manuscripta.se, a catalog of Medieval and Early Modern manuscripts in Sweden.[6] This interview focussed on some of the concrete problems Patrik faced while developing that resource. Although to some extent anecdotal, this discussion will serve to highlight the impact the tailored digitization process can have on the resulting digital facsimile images, and how their description in the form of paradata might benefit the desired end product (be that called an archive, catalog, or edition).

As is usual for a project such as manuscripta.se, it was developed in a series of phases that depend largely on research funding cycles. In its initial phase, the project took place at the Uppsala University Library from 2012 to 2016 and aimed to construct a new, improved, digital catalog of all 130 Greek manuscripts in Sweden.[7] Most of these manuscripts are held at the University Library in Uppsala, but manuscripts were found in holding libraries across Sweden. When I interviewed Patrik, the project was in its second phase, having moved with him to KB-SE. The project's goal at the time was to expand the original catalog with 276 more medieval manuscripts in Sweden that were held either in the Uppsala University

[5] See: https://dixit.uni-koeln.de.

[6] See: https://www.manuscripta.se/.

[7] For more details, a final report on the project can be found in Nyrström (2016).

Library or in KB-SE itself. These manuscripts were written in Old Swedish.[8] At the time of writing, the project has completed this work and has moved into a new phase, where KB-SE's Latin manuscripts are added to the catalog.[9] Our interview focussed mostly on Patrik's experiences in the project's first phase, because at the time that phase of the project incorporated its most significant digitization effort (because it was the only phase in the project that had allocated a budget specifically to the digitization of the manuscripts in question).

As the first phase of the project was hosted by the Uppsala University Library, the library that had also collected most of the Greek manuscripts in Sweden, the bulk of the materials were digitized by this institution. Besides the Greek manuscripts in its own holdings, manuscripts that are held by other Swedish university libraries (specifically Gothenburg University Library, the Linköping Diocesan Library and the Skokloster Castle Library) were transported to Uppsala to be digitized there, by the university library's designated photographer. However, some memory institutions that have their own specialized digitization services (such as KB-SE and Riksarkivet—the Swedish national archive) insisted on digitizing the manuscripts themselves rather than transporting them to Uppsala. This invariably led to differences in approaches and standards. In Uppsala for instance, being a local rather than a national memory institution, digitization was still in its early stages. This gave Patrik the chance to provide more input for the development of the institution's digitization workflow, and to place certain demands on the quality of the reproductions. At KB-SE, on the other hand, digitization practices were more advanced and noticeably more professional, and while this may have been beneficial for the initial quality of the reproductions, it also gave Patrik less of a chance to influence the digitization process. At Riksarkivet, then, Patrik had even less opportunity to place demands on the quality of the reproductions, as its high-end tier of digitization services was too expensive for the project, obliging the team to opt for lower-resolution scans of the manuscripts. This means that there is considerable qualitative variance for the reproductions that were produced at different memory institutions throughout the project.

Besides these differences in quality between institutions, Patrik also suggested that there may be considerable differences within the selection of reproductions produced by the Uppsala University Library—partly *because* he had more of an influence on the digitization practice. As the institution did not have much experience with digitization at the time the project started, some improvements

[8] A final report on this project (in Swedish) can be found in Nordin and Ahlbom (2021).

[9] For more information, see: https://www.rj.se/en/grants/2021/medieval-latin-manuscripts-in-the-national-library-cataloguing-and-digitization/. Alongside this current phase of the project, manuscripta.se is also developed further through the "Swedish Post-medieval Manuscripts at the National Library of Sweden and Uppsala University Library" project (a spin-off project of phase 2; for more information see: https://www.rj.se/en/grants/2018/swedish-post-medieval-manuscripts-at-the-national-library-of-sweden-and-uppsala-university-library--a-catalouging-and-digitization-project/), and a related project on West Norse manuscripts (see: https://www.rj.se/en/grants/2021/digitization-of-the-west-norse-manuscripts-in-swedish-collections/).

were made during the course of the project, which were not always retroactively corrected. The most important factor here is perhaps the improvement of the institution's equipment. At the start of the project, the library acquired a new book scanner for the specific purpose of this digitization project. After producing a few test batches, however, it became clear that the quality of the reproductions did not suffice for the project. Asked for examples, Patrik explained that the resolution was too low, the colors were inaccurate, and that the scanner's dual light source produced a flare in the middle of the scan. Confronting the photographer with this problem, it was agreed that the quality of the reproductions needed to be improved and that a new solution had to be found. This is a clear example of how a researcher's demands may influence the negotiation of the quality standards of digitization practices at memory institutions. Patrik explained that he pushed hard to have the manuscripts digitized with a camera, which in the end they were: the test batch was discarded and reshot with a 20 megapixel Hasselblad camera. In the course of the project, however, the institution made a new investment in its equipment and purchased a 60 megapixel camera, which, from that point onward, was used for the project instead. The result is that approximately $\frac{1}{3}$ of the reproductions at the Uppsala University Library were digitized at a lower resolution.

Another example of an improvement to the digitization practice at Uppsala University Library that was not retroactively corrected was a better control of the lighting conditions in the designated photography room. Patrik explained that in the beginning of the project, the door of this room was not always closed, allowing some natural light to come into the room, which affected the lighting and color of the reproductions. This problem becomes especially apparent when you start compiling spreads—placing each recto alongside its rightful verso—which allows for a close comparison of facing leaves. If the two photographs were shot in exactly the same lighting conditions, there should be no color difference between recto and verso, allowing them to be stitched together seamlessly. But since rectos and versos were photographed consecutively (a common practice in heritage photography to speed up the workflow), and natural light was allowed to influence the digitization process, this could lead to a distinct color variance between recto and verso in individual spreads. In the course of the project, this problem was discovered and eventually solved, but not retroactively corrected, leading to some differences in the quality between individual reproductions digitized at Uppsala University Library, even within the same manuscript.

These instances suggest that there was a good dialogue between the researcher and the photographer, allowing both parties to learn a lot from each other in the process. The advantage of such close collaboration at a point where a memory institution is still developing its digitization practices was that it allowed Patrik to help steer the construction of the workflow to some extent, giving him more control over the final state and quality of the reproductions. The disadvantage, on the other hand, was that since the quality standards of the reproductions were still evolving over time, this produced a discrepancy in the quality of early versus late digitized reproductions. And this evolution continued on to the time of the interview: confident that this collaboration had taught both himself and the institution's photographer a lot and brought them more or less on the same page with regard to quality control issues, Patrik nevertheless felt that there was still room for improvement and that a continued close collaboration between both parties would be advisable and mutually beneficial.

At an institution with a developed digitization workflow and carefully constructed quality standards like KB-SE, researchers may not be able to influence the reproduction of the originals as much, but there would arguably be less reason to do so in the first place. Although Patrik suggested that he would still think it necessary to check the quality of each of the images himself, the fact that KB-SE has a longstanding tradition of developing guidelines and setting (high) standards for their digitization practices inspires more confidence in the resulting reproductions and would hopefully decrease the time spent discussing issues with the photographers and reshooting individual pages as a result. In addition, the fact that a memory institution has an established digitization workflow that is less prone to change can also be an advantage. Especially for high-volume digitization projects like manuscripta.se, this would make the quality of the reproductions more consistent, and therefore also more easily accounted for. Even in the case of low-quality images, documented digitization guidelines and workflows at least give the researcher a tangible point of reference to explain any discrepancies between the original and the reproduction to the user. In the end, this is where the researcher will have to decide what still constitutes an acceptable quality standard for the edition, and to what extent it is worth to surrender image quality of the reproductions in order to secure an improved accountability for the edition.

This brings us to the final stage of the manuscripta.se project: to present the reproductions of these manuscripts as part of the digital catalog the team was developing. To do so, the catalog would need an interface—a publication platform that would make the digital information accessible to the user. The software that was used to display the images on the website is called IIPImage. To work, this software requires the images to be saved on the server in a multilayered TIFF or JPEG2000 format. For manuscripta.se, Patrik chose to work with the first option. To achieve this format, Patrik needed to process the (regular) TIFF images he received from the library, to turn them into tiled "Pyramidal TIFF" images before putting them on the server. Patrik reported that he did not perform any additional postprocessing on the images (except changing their file names to also include their folio number). He suspected that the photographer did perform some postprocessing to the presentation copies of the digital reproductions, such as cropping, color adjustment, and sharpening—a practice that would be consistent with KB-SE's, but these were not explicitly communicated to him. This is a good example of some aspects of the decision-making process that remained invisible—even to the person who requested the images.

In the delivered images, a standard margin was left around each manuscript page, and a ruler was photographed in this margin to attest to the page's size. The images were not rescaled to 1:1 at 300ppi (as is KB-SE's default practice) because the manuscripts vary in size (sometimes being as small as 10 cm wide), which would have rendered text on some of the smaller manuscripts illegible.

As discussed above, Patrik checked the quality of the images extensively before uploading them to the server, comparing them to the manuscript (leafing through the original while inspecting the digital copy). Sometimes color correctness, brightness, or focus issues were discovered, but the most frequent errors would be missing pages and the fact that a bookmark or part of the manuscript was concealing some of the text. In these cases requests were made to shoot or reshoot the relevant pages. When Patrik was satisfied with the quality of the digital reproductions, he uploaded them to the server and made them accessible to his users.

This case study shows that, depending on the memory institution, the researcher who commissions the digitization of specific cultural heritage objects may have more or less of an influence on the digitization practice and illustrates what the advantages and disadvantages of such influence may be. It also highlights the kind of demands that the researcher will place on the digital reproductions and how the negotiation of quality standards between researcher and photographer may take place. In Patrik's case, there was already a difference here in his dealings with the various memory institutions. And indeed, for a large part, Patrik found himself in quite a rare position, where the connections he had established earlier with the library staff allowed him to be more closely involved in the process. To a less connected patron, such a privileged position may not be available, and even more parts of this process may remain invisible to them as a result. The example also indicates that this digitization still happens on an ad hoc basis and that digitization practices across memory institutions have not yet been standardized. This means that there will be considerable differences in quality and standards across memory institutions—or even within a single memory institution as guidelines and workflows are still under development, and its equipment and staff undergo changes. It should also be noted that at least in this particular case, the researcher was more concerned with getting the best possible quality for each individual image (pushing its quality as far as possible depending on the rapport he had with the photographer, and the type of quality the project's budget could feasibly afford) rather than with providing a more consistent overall quality for all the images, based on a single and more easily documented quality standard.

At the moment of writing, the catalog's reproductions of the manuscripts are offered without any reference to the quality standards that were used to photograph the images, but Patrik attested (based on his interactions with the users of his catalog) that the lack of such a description is also not an issue that the user is particularly concerned with. In part, this can be explained by the fact that manuscripta.se is a catalog first—not an image repository, nor a digital scholarly edition. Still, even in the case of scholarly editions, attesting to the way in which the digital reproductions that are used in the edition were produced is not a common practice. But that is of course not a reason why we should not start doing so. Indeed, as these digital reproductions are becoming more and more detailed and important in our digital scholarly editions, we should also become more and more vigilant about the way we present them (and with it: their relation to the original) to our users.

4 Paradata for Digitization Processes

As became clear in my comparative study of National Libraries for the DiXiT project, these high-level memory institutions usually offer several tiers of digitization services, from low-end services (with the help of librarians and the library's most basic equipment) to high-end services (performed by photographers with professional equipment). Which of these services are relevant and available for a specific project is evaluated on a case-by-case basis, in a process of negotiation between the library's staff and its patrons, that depends on a wide range of factors—such as the material state of the document, the needs of the project, the availability of equipment, the experience of the staff, and the funding that is available.[10] As my case study has demonstrated, the fact that so many aspects of this process potentially remain invisible to the end user is problematic, from a scientific perspective, because they can play an important role in the project. For example, the equipment and professional expertise a library has at its disposal may change over time, and certain tiers of digitization services may be unavailable through a lack of funding. Intuitively, one would think that a given memory institution's high-end digitization services would be the ideal environment to procure the high quality, consistent digitization results that a prestigious high-quality digital scholarly edition (or catalog) requires—not just because it is potentially the most qualitative option (using the best equipment, producing the highest resolution scans and truest color reproductions), but also because it is usually the only option that follows high predetermined standards, where the digitization is performed by trained professionals, and takes place in a more or less controlled environment.

As the above case study has demonstrated, however, digitization is not a purely mechanical process that produces identical results at every iteration. To the contrary, it is a highly personal process that involves a number of different agents, each with their own experiences and demands, who are at times quite literally involved in a process of negotiation between themselves (based on personal preferences and professional experiences), the physical limitations of the document, the environment where the digitization takes place, and any contractual obligations that need to be met—all of which have a direct impact on the quality and consistency of the digitized end product. Indeed, as we have seen, there is still plenty of room for variation in quality among the delivered image files—even over the course of a single digitization project, conducted by a single team, at a single institution, performing a single digitization service. In some cases (e.g., when overall consistency is more important than factors such as image resolution) this may mean that a more

[10] I performed a more detailed analysis of these different tiers at KB-SE and proposed a mapping of the negotiations between different internal and external agents in the digitization process as part of my work on the DiXiT project. Because of space restrictions, however, it was impossible to include those findings in this chapter. I did, however, already discuss some of these aspects in a blog post (Dillen, 2017a).

mechanical, lower-end digitization service (like an automated book scanner) might be preferred over high-end alternatives.[11]

Whichever suitable service or digitization process is agreed upon, however, the case study highlights how important it is to provide some sort of context for the images you use (such as their origin and some specifications of the digitization process) to justify their quality (or lack thereof). This is especially crucial because a lower quality invariably points to a greater divide between the original and the reproduction and, in the case of a scholarly edition, potentially between the reproduction (which is often the user's only point of reference) and the transcribed text. This divide needs to be addressed and accounted for somehow before the editor's transcriptions can really be considered "assessable" by the user—which is arguably the whole point of adding these images to the digital scholarly edition in the first place. This is where paradata could enter the picture.

Paraphrasing my definition of paradata at the start of this chapter, we could say that it concerns data that describes data creation processes. In some cases, paradata comes in the form of metadata (highly structured "data about data," described in metadata terms, following a specific ontology like DCMI,[12] or CIDOC-CRM[13]—to stick with examples that are relevant for memory institutions). For an example of what this may look like for digitization practices, we can consider EXIF (Exchangeable Image File Format)—a standard for storing metadata in image and audio files.[14] Nowadays, most digital cameras use this format to automatically store information about the equipment and settings that were used directly in the image file, at the moment the image was captured. This may include the make and model of the camera or scanner that was used, the model of its lens, together with the aperture and shutter speed that were used, etc. All of this data is structured as metadata and is highly relevant to the process of the image's creation process, which means it qualifies as paradata too. But not all metadata is paradata. Say, for example, that we have a book in the collection of a library. And that the book's metadata includes a Dewey Decimal Class (DDC) number, which classifies the book as belonging to a particular discipline and helps the library's patrons locate the document. This number is part of the book's metadata because it is a highly structured record that tells us something more about the book (i.e., it is "data about data"). But it is not necessarily paradata, because this number was given to the book by a librarian, in retrospect, based on their interpretation of the document—and tells us nothing about the circumstances in which the document was originally created. At the same time, not all paradata is metadata. Take, for example, the interview I conducted with Patrik. This interview is a source of data in itself, recorded as an audio recording and

[11] This is true, for example, in cases where the editor aims to publish a lightweight minimal edition in order to reach a wider, less privileged audience; or when the editor has a limited budget and needs to put the text first.

[12] See: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[13] See: https://cidoc-crm.org/html/cidoc_crm_v7.1.1.html.

[14] See: https://exiftool.org/TagNames/EXIF.html.

accompanying transcript, which has served to support some of the arguments I made in this chapter. And it qualifies as paradata, because it gives an account of a set of conditions and practices that had a direct impact on the creation of the data files (in this case: images) that would later become part of the manuscripta.se database. But it is not strictly metadata because it is not structured and mapped onto an existing ontology of metadata terms.

What this means is that paradata is very much in the eye of the beholder and that what does or does not count as paradata in a given context greatly depends on the phenomenon that is being studied. So we would do well to specify exactly which data creation workflows and processes we are dealing with, before considering what the paradata around these processes may look like, and what the benefit would be of sharing them with the research community. In turn, this will help us "explicat[e] in depth what to document and how to capture it" (Huvila, 2022, p. 33).

In the context of this chapter, the process we are interested in is the digitization process. Specifically, we are looking into the way in which we create digital reproductions of physical objects. With regard to this digitization process, we are not just focussing on the moment of capture (e.g., when the digital photograph is taken, the scan completed, or the video recorded), but also (and: especially) on the workflows that dictate how this act of capturing is performed, and how the resulting reproduction is processed and evaluated afterward. And as we have seen, these workflows are basically decision-making processes where different agents interact with each other, and directly or indirectly negotiate how the physical documents and their digital reproductions are handled—agents who each put their own demands on the physical and digital objects concerned, based on a strong foundation of professional expertise. This means that the primary sources of our paradata are human agents and that the most straightforward way of obtaining such paradata is through qualitative social science research methods, such as interviews, surveys, observation studies, etc. In addition, these primary research data could potentially be complemented with secondary materials, such as institutional web pages, internal (digitization) policy documents, descriptions of workflows and standards, third-party contracts and terms of agreement, automatically generated metadata, changelogs, paper trails of internal and external communication, etc. Combined, all these materials can help paint a detailed picture of exactly how the physical object was digitized, and why it was digitized in this way.

5 Paradata for Digital Scholarly Editions

In this last section of our chapter, I would like to reflect on what these findings mean for the digital scholarly edition. We would be too hasty to conclude, for example, that from this point onward, any digital scholarly edition worth its salt needs to incorporate all the relevant paradata pertaining to the images it uses, from EXIF metadata to research diaries and recorded interviews with any agent who was involved in the digitization process in some way or other. While a close examination of digitization practices at a given memory institution with a view to developing a portfolio of relevant paradata may be a worthwhile research project in itself that can provide us with new, valuable insights about the exact relation between physical objects and their digital reproductions, requiring such an investigation to be attached to each and every digital scholarly editing project is simply infeasible and arguably lies well outside the purview of the editorial endeavor.

To some extent, this is due to the nature of the paradata involved. Some paradata is easy to come by, but extremely hard to understand. For example, a data dump of metadata, technical specifications, and changelogs about the images that were used in the digital scholarly edition would be relatively easy to extract and provide alongside the digital scholarly edition. But without additional context, such data would be lost on the average user of the edition, who is mainly interested in the edition's documents and their interpretation, and who does not necessarily know how to read (let alone assess) this kind of technical documentation—thereby defeating the purpose of the paradata. On the other hand, some paradata is easy to understand, but extremely hard to come by. Providing detailed descriptions of digitization workflows, interactions, and decision-making processes on the basis of interviews, observation studies, and other social science methods is simply infeasible for most digital scholarly editions. They would add a whole new line of inquiry to the research, which would require the team of editors to add an entirely new research profile to their project, with its own allocated researchers, time, and budget. This would put even more of a burden on the editorial team than is already the case.[15] Requiring this kind of treatment of any self-respecting digital scholarly edition would turn it even more into a prohibitively expensive research endeavor than it already is.

Still, it may be possible to find a middle way. As Huvila remarks in the concluding paragraphs of his essay: "an equally critical question to having enough [paradata] is how to avoid having too much" (Huvila, 2022, p. 41). And indeed, since different data users have different data needs in different contexts and situations, it is essential that we first try to understand which types of paradata are "useful and usable" in our particular context (Huvila, 2022, p. 29). In our case, this context is that of a scholarly editor developing a digital scholarly edition on the basis of a (series of) document(s) held by one or more memory institutions. In this sense, the scholarly editors are not exactly the creators of digital reproductions of physical cultural heritage documents, but rather the reusers of those reproductions.

[15] In her seminal work *Digital Scholarly Editions: Theories, Models, Methods*, Elena Pierazzo exposes how the digital turn has essentially required the editor (or editorial team) to acquire an entirely new set of skills. Where in the age of print the scholarly editor would mainly require a deep knowledge of philological skills such as "textual scholarship, codicology, palaeography, historical linguistics, literary criticism," etc., moving the discipline into the digital age requires a whole new set of computational skills, such as data management, web design, digital infrastructure, tool development, and metadata standards (Pierazzo, 2015, p. 126). If we would require all the relevant paradata related to the digitization process of all relevant digital reproductions to be included in the edition as well, this would expand the list of skills even more, to include an understanding of social science research methods, digital photography, library collection management, etc.

This means that while paradata that attests to the way these digital reproductions are created at a given memory institution would be relevant information to the scholarly editor (as well as to the resulting digital scholarly edition's user), it is not the editor's responsibility to commission, store, and distribute those data in the first place. Instead, the scholarly editor is reusing the memory institution's data (be they physical, digital, or both) in order to make their own digital object: the digital scholarly edition. And it is paradata detailing how the latter digital object was created that contains essential contextual information for the digital scholarly edition's (re)users.

The practice of describing exactly how scholarly editions are made is not new to the field of scholarly editing. In fact, it is arguably one of its strongest foundations. If centuries of textual critical thinking have taught us anything, it is that the authority of some documents, texts, or versions over others is very much a matter of interpretation, and that opinions differ greatly when it comes to, for example, the extent to which the scholarly editor is allowed to intervene in the texts they are editing. And it is exactly this realization of how much the presentation of the scholarly edition ultimately depends on the editor's perspective that has since long required the scholarly editor to position their research in the field and justify any editorial decisions they have made in a designated section that is appended to the scholarly edition.[16] This section, which has existed long before the discipline took a digital turn, is often called "Editorial Principles" or something similar and does exactly this: explicate the editorial principles on which the edition is based, and the rules the editorial team have followed to construct the edition, with as few exceptions as possible. Since the digital turn, this section has only gained more significance and has been expanded to also include some type of "Encoding Description" and other "Technical Documentation." In these sections the editors do not just explain the theoretical framework they have used, but also which technical standards they have used,[17] and which choices they have made within those standards, to put this framework into practice. Such sections have become a staple of the (digital) scholarly edition—to the extent that they are, for example, repeatedly mentioned in the evaluation criteria used by RIDE, the well-respected review journal for digital editions and resources in the field (Sahle et al., 2014).

These sections can be regarded as important sources of paradata *avant la lettre*. They are indeed narratives that embed useful information about the way in which a digital dataset was created—in our case: the digital scholarly edition. As scholarly editors have realized a long time ago, this type of information is absolutely essential if we want our data (be it digital or in print) to be used, understood, and potentially reused. And it is exactly by enclosing this information in the edition that it can

[16] The parallel development of various schools of textual scholarship is a well-established part of the research field's history that I explored in some more detail in my PhD thesis (Dillen, 2015). For more information and a more practical resource, see the Lexicon of Scholarly Editing (Dillen, 2020).

[17] Such as, for example, the guidelines of the Text Encoding Initiative (https://tei-c.org).

achieve one of its most important goals: to inspire a critical attitude toward texts in general, and the relevant work(s) in particular. What seems to be missing from these sections, however, from the perspective of this chapter, is some similar information that helps contextualize the authority of the *images* that were used in the dataset. This information does not necessarily need to be as detailed as all the paradata we can imagine to contextualize the digitization process—as described above. The important information pertaining to the relevance and reusability of the images that were used in the digital scholarly edition is not so much exactly how the digital surrogate resembles or differs from the original document, but rather the extent to which the editor recognizes that the digital reproduction are sufficiently representative for the context in which they are used, despite its specific limitations. Ultimately, it is the editor who vouches for the authority of the images that are used in the edition, in the same way that they vouch for the authority of their texts. And like with their text—the authority of which is justified in dedicated sections detailing the edition's "Editorial Principles," the editor would do well to justify the authority of the images that were used by providing more context relating the way in which they were obtained. In both cases, these justifications serve to convince the reader of the soundness of the editor's research and argumentation, or, failing that, to point the reader in the direction of the originals whenever they doubt the editor's claims.

Writing such a section should not take up a disproportionate amount of time and space, nor should it require the editor to acquire any additional skills. What is needed is simply some additional context that frames the editor's decision-making process around the time when they made or acquired the digital images, and determined to what extent they were sufficiently representative of their originals for the digital scholarly edition. This could be achieved, for example, by describing their own role in the digitization process—however large or small this may have been. This may include a brief description of the specific digitization service they engaged (and, where relevant, whether this choice was the result of a process of negotiation), which requirements and technical specifications they requested from the digitization team (and, where relevant, whether these requirements were met), whether or not the editor established a rapport with the digitization team (and, when relevant, whether this feedback loop had any effects on the overall quality of the images), and which measures the editor took to ensure quality and representativeness of the reproductions, in relation to their originals (for example, by comparing the reproductions to the originals to make sure they are accurate and complete). By doing so, the editor would acknowledge their position as one agent in a complex process while, at the same time, explicitly taking responsibility for this crucial aspect of the digital scholarly edition. This would provide the reader with the necessary information to decide whether or not to trust the editor's judgment, thereby drastically improving the accountability of the edition. At the same time, such a section would raise the general awareness of the complexity of digitization processes and acknowledge that the resulting digital reproductions are objects in their own right, distinct from their physical originals, and differing from them in significant ways. All of this can only help inspire a critical attitude in the reader and open their mind to new interpretations of treasured cultural heritage documents. And that, to a large extent, is exactly what the fields of textual scholarship and scholarly editing are all about.

Acknowledgments As the author of this essay, I would like to thank Patrik Granholm for agreeing to what would turn out to be a very informative interview, and for granting me permission to discuss the contents of that interview in this publication. The research leading up to these results was funded in part by the Digital Scholarly Editions Innovative Training Network (DiXiT ITN), a Marie Sklodowska-Curie Action, underwritten by the EU Seventh Framework Programme (FP7/2007–2013), REA grant no. 317436.

References

Andrews, T. L., & van Zundert, J. J. (2018). What are you trying to say? The interface as an integral element of argument. In *Digital scholarly editions as interfaces* (pp. 3–34).

Apollon, D., Bélisle, C., & Régnier, P. (Eds.) (2014). *Digital critical editions*. University of Illinois Press.

Björk, L. (2015). *How Reproductive is a Reproduction? Digital Transmission of Text-based Documents*. PhD Thesis, University of Borås, Borås, 2015. http://www.diva-portal.org/smash/get/diva2:860844/INSIDE01.pdf

Bleeker, E., & Kelly, A. (2018). Interfacing literary genesis. In R. Bleier, M. Bürgermeister, H. W. Klug, F. Neuber, & G. Schneider (Eds.), *Digital scholary editions as interfaces* (Vol. 12, pp. 193–218). BoD–Books on Demand.

Boot, P., Cappellotto, A., Dillen, W., Fischer, F., Kelly, A., Mertgens, A., Sichani, A.-M., Spadini, E., & Van Hulle, D. (Eds.) (2017a). *Advances in digital scholarly editing: Papers presented at the DiXiT conferences in The Hague, Cologne, and Antwerp*. Sidestone Press. ISBN 978-90-8890-483-7. https://www.sidestone.com/books/advances-in-digital-scholarly-editing.

Boot, P., Fischer, F., & Van Hulle, D. (2017b). Introduction. In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, & A. Kelly (Eds.), *Advances in digital scholarly editing: Papers presented at the DiXiT conferences in the Hague, Cologne, and Antwerp* (pp. 15–22). Sidestone Press. https://www.sidestone.com/books/advances-in-digital-scholarly-editing

Dillen, W. (2015). *Digital Scholarly Editing for the Genetic Orientation: The Making of a Genetic Edition of Samuel Beckett's Works*. PhD Thesis, Universiteit Antwerpen, Faculteit Letteren en Wijsbegeerte, Departement Letterkunde. https://hdl.handle.net/10067/1305290151162165141.

Dillen, W. (2017a). ER1 - Reminiscing on a Year of DiXiT. https://dixit.hypotheses.org/1384.

Dillen, W. (2017b). What you c(apture) is what you get: Authenticity and quality control in digitization practices. In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. Van Hulle (Eds.), *Advances in digital scholarly editing: papers presented at the DiXiT conferences in the Hague, Cologne, and Antwerp* (pp. 397–400). Sidestone Press. ISBN 978-90-8890-483-7. https://www.sidestone.com/books/advances-in-digital-scholarly-editing

Dillen, W. (2018). The editor in the interface: Guiding the user through texts and images. In R. Bleier, M. Bürgermeister, H. W. Klug, F. Neuber, & G. Schneider (Eds.), *Scholarly digital editions as interfaces*. Number 12 in SIDE: Schriften des Instituts für Dokumentologie und Editorik (pp. 35–59). Books on Demand. ISBN 978-3-7481-0925-9. http://kups.ub.uni-koeln.de/9111/

Dillen, W. (2020). *A Lexicon of Scholarly Editing*. https://doi.org/10.5281/ZENODO.4008433

Dillen, W. (2022). *What is a Digital Scholarly Edition?* https://play.hb.se/media/0_v96oltn7

Doran, M., Edmond, J., & Nugent-Folan, G. (2019). Reconciling the cultural complexity of research data: Can we make data interdisciplinary without hiding disciplinary knowledge? *Preprint of manuscript submitted to CODATA*.

Driscoll, M. J., & Pierazzo, E. (2016). *Digital scholarly editing: Theories and practices* (Vol. 4). Open Book Publishers. https://books.openbookpublishers.com/10.11647/obp.0095.pdf

Franzini, G., Terras, M., & Mahony, S. (2016). A catalogue of digital editions. In M. J. Driscoll, & E. Pierazzo (Eds.), *Web.* https://github.com/gfranzini/digEds_cat (pp. 161–182). Open Book Publishers. https://www.openbookpublishers.com/reader/483/#page/180/mode/2up

Huvila, I. (2022). Improving the usefulness of research data with better paradata. *Open Information Science, 6*(1), 28–48. ISSN 2451-1781. https://doi.org/10.1515/opis-2022-0129 https://www.degruyter.com/document/doi/10.1515/opis-2022-0129/html.

Nordin, J., & Ahlbom, K. (2021). *TTT: Text till tiden! Medeltida texter i kontext – Då och nu. Slutrapport*. Technical Report, Kungliga Biblioteket, Stockholm. https://lucris.lub.lu.se/ws/portalfiles/portal/103218558/TTT_slutrapport_210622.pdf

Nyrström, E. (2016). *Greek manuscripts in Sweden - a digitization and cataloguing project*. https://www.rj.se/en/grants/2011/greek-manuscripts-in-sweden---a-digitization-and-cataloguing-project/

Pierazzo, E. (2015). *Digital scholarly editing: Theories, models and methods*. Ashgate Publishing Limited.

Sahle, P. (2008). *A Catalog of Digital Scholarly Editions*. https://www.digitale-edition.de/exist/apps/editions-browser/\protect\T1\textdollarapp/index.html

Sahle, P. (2013). *Digitale Editionsformen*. SIDE: Schriften des Instituts für Dokumentologie und Editorik (Vol. 7–9). BoD–Books on Demand.

Sahle, P., Vogeler, G., Broughton, M., Cummings, J., Fischer, F., Steinkrüger, P., & Scholger, W. (2014). *Criteria for Reviewing Scholarly Digital Editions, version 1.1* |*, 2014*. https://www.i-d-e.de/publikationen/weitereschriften/criteria-version-1-1/

Zeller, H. (1971). Befund und Deutung. Interpretation und Dokumentation als Ziel und Methode der Edition. In H. Zeller & G. Martens (Eds.), *Texte und Varianten. Probleme ihrer Edition und Interpretation* (pp. 45–89). CH Beck, München.

Wout Dillen is a Senior Lecturer in Library and Information Science at the University of Borås. He holds a PhD in Literature with a focus on text encoding and digital scholarly editing from the University of Antwerp. From 2016 to 2017, he held a Marie Skłodowska-Curie Experienced Research Fellowship in the Digital Scholarly Editions Initial Training Network (DiXiT ITN) at the University of Borås, Sweden. Wout is the Secretary of the European Society for Textual Scholarship (ESTS) and General Editor of its journal Variants. He is also an Executive Board member of DH Benelux and a coeditor of its journal.

Reconstructing Provenance in Long-Lived Data Systems: The Challenge of Paradata Capture in Memory Institution Collection Databases

Alexandria J. Rayburn and Andrea K. Thomer

Abstract

Paradata is important for understanding the provenance of data—but capturing and using paradata is challenging because it is often not formalized or explicit. This is particularly the case for complex, long-lived digital objects, such as the databases used to manage long-lived museum collections. These databases are passed down between generations of collections managers, but the documentation explaining their structure and changes over time is often incomplete, thus posing an obstacle to the use and maintenance of the databases. Collection managers must often reverse engineer their databases and create documentation from scratch. Here, we present a case study of paradata reconstruction conducted as part of a larger project studying database maintenance in memory institutions. Through interviews with collection managers at the University of Michigan Herbarium and Matthaei Botanical Gardens, we reconstruct how a database evolved and changed over 50 years. We show how different ways of illustrating the history of a database can be used to help "open up" a database for users. We reflect on the strengths and weaknesses of these approaches, specifically versioned entity relationship diagrams, Sankey diagrams, and narrative case summaries, and discuss the challenges in capturing paradata from long-lived sociotechnical objects.

A. J. Rayburn (✉)
School of Information, The University of Michigan, Ann Arbor, MI, USA
e-mail: arayburn@umich.edu

A. K. Thomer
School of Information, The University of Arizona, Tucson, AZ, USA
e-mail: athomer@arizona.edu

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_9

1 Introduction

Literary theorist Gérard Genette famously described paratext (e.g., the table of contents, chapter headings, index, and other framing text in a book) as "a threshold, or… a vestibule" and a "means by which a text makes a book of itself and proposes itself as such to its readers" (Genette, 1991, p 261). Para*data*, however, is typically described in less poetic terms: it is a form of metadata that describes the ways in which a dataset was collected, processed, or manipulated (Pomerantz, 2015). This nuts-and-bolts definition elides paradata's role in similarly presenting data and digital objects for use and interpretation. Paradata, like paratext, is a threshold through which users encounter, work with, and manage complex digital objects. It presents and preserves aspects of a dataset's context of production and fundamentally shapes how an object is viewed. And when paradata is missing or incomplete, datasets and digital objects become harder—if not impossible—to "enter" or otherwise engage with.

In this chapter, we present a case study that illustrates the challenges of working with a particularly complex, long-lived digital object—a museum collection database—without complete documentation of its development or change over time. Museum collection databases are used to store information about the artifacts, specimens, and other objects in a museum's collection. For many museums, their collections date back decades if not centuries, and their associated databases are decades old as well. The stewardship of these databases is "passed down" from one collection manager to the next, but, for a variety of reasons, sometimes without significant documentation. This leaves the "new" collection manager with the unenviable task of reverse engineering a database's structure, contents, and data entry workflows for the next generation (Thomer et al., 2018; Thomer & Rayburn, 2023). Even once reverse engineered, not knowing the reasons underlying database design decisions makes database usage difficult.

Our case study focuses on the evolution of two collection databases with a common origin: the Matthaei Botanical Gardens and Nichols Arboretum (MBGNA) database and the University of Michigan (U-M) Herbarium collection database. Both databases grew out of a system called TAXIR in the 1960s; both have been repeatedly migrated between different software systems over the years; and both lack significant documentation explaining their origin, structure, or evolution (leading to much frustration for both databases' current collection managers). As part of a larger project studying memory institution database maintenance over time, we interviewed database stewards at each museum and "read" each site's databases (Feinberg, 2017) to reconstruct change in each site's databases. We also tried different approaches of illustrating this change over time: first, versioned entity relationship diagrams (an illustration of a database's underlying data model), and second, Sankey diagrams (an illustration of the "flow" of records between different database versions). In retrospectively creating this documentation, we ask: *how and to what degree can we record the sometimes-subtle changes that occur in a database over long periods of time?*

Our chapter proceeds as follows: we first contextualize our project in prior critical scholarship on databases, their paradata, and their use in memory institutions. Then we present our case study and the diagramming techniques we used to reconstruct database histories. We conclude by reflecting on the strengths and weaknesses of these approaches. We find that when paired together, entity relationship diagrams, Sankey diagrams, and narrative histories can provide new users an entry point into these complex data systems.

2 Background

Most fundamentally, a database is a structured collection of data organized for fast search and retrieval by a computer (Manovich, 2002). More abstractly, databases are tools for encoding the world (Dourish, 2017); that is, they translate non-computational entities into a machine-readable form. Database technology has drastically expanded possibilities for organizing complex or difficult-to-collect data and has played a critical role in allowing research data to be shared between collaborators, or in building networks for data sharing (Bruns et al., 1998; Cullings & Vogler, 1998; Mineta & Gojobori, 2016; Robertson et al., 2014; Stein & Wieczorek, 2004; Vieglais et al., 2000; Williams et al., 2018). There are many different technical standards for databases; here we are focusing on relational database management systems (RDMS), in which data is stored in a series of linked tables, each table connected through data points. The structure of these tables (the data model) is customized based on the nature of the data or its relevant metadata.

Databases, like most digital artifacts, are quite fragile and require constant maintenance to persist. They are sociotechnical objects, built on ever-changing physical and digital infrastructures and fundamentally shaped by the people and organizations that create them (Bowker & Star, 2000; Dourish, 2017; Hine, 2006). Different aspects of a database change at different rates, creating significant challenges for the maintainers of these systems. For example, in memory institutions—the museums, archives, and libraries that “contain memory of peoples, communities, institutions and individuals” (Dempsey, 2000)—the database’s structure and hardware might not change very often, but the users and needs of users of that database change fairly frequently (Fig. 1). This increases the chances that a system will break or require alteration or repair over time. As different parts of a database become obsolete over time, the database will need to be migrated from one piece of hardware or software to another. As we explore later, this can quickly become a complex process as migrations are often catalysts for other structural changes to the data, such as a schema change.

Critical to database migration is an understanding of the systems’ accompanying paradata. Though the term paradata was originally used by statisticians to describe process data needed to understand the quality of statistical data, such as surveys (Karr, 2010; Kreuter et al., 2010), the term has since come to be applied beyond survey data. However, as the term has gained adoption, it has also become more challenging to define. In many domains, paradata is often left undocumented or

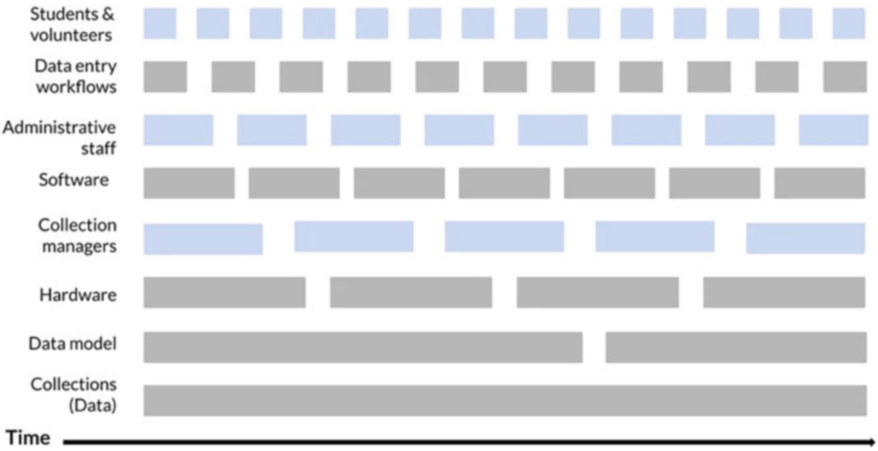


Fig. 1 A model for sociotechnical change in museum collections. Note the varied rates that change occurs, each factor contributing to data migrations and its accompanying paradata. Reprinted from Thomer & Rayburn, 2023 with permission

implicitly present within the data itself (rather than explicitly recorded) (Börjesson, 2022; Huvila et al., 2021). Paradata can take many different forms depending on the nature of the data it contextualizes. Identifying the form that paradata will take within a certain context is an endeavor on its own. Indeed, much literature examines the forms of paradata as a research question. For example, Börjesson et al. (2022) discuss the different forms of paradata that appear within archeological field data, while West and Sinibaldi (2013) examine types of paradata within social science datasets.

In this case study, the paradata we are speaking of is any provenance or processing data related to database creation, upkeep, and migration of natural history museum data. Natural history collections often contain millions of specimens that date back hundreds, even millions, of years, with each specimen requiring documentation. These data-intensive collections were early adopters of database technology and early contributors to scholarship in data curation (Palmer et al., 2013; Strasser, 2012; Thomer et al., 2018). Because of these long data histories, the paradata associated with these collections is lengthy and complex. Further, the users, creators, and maintainers of memory institution databases have varying work practices, often very different from best practices defined in computer science (Thomer & Wickett, 2020). These idiosyncratic practices are where much paradata is created. Though there are systems designed to track the provenance and changes to databases over time (Brodie & Stonebaker, 2015; Buneman et al., 2006, 2009), most of the databases in natural history museums (NHM) do not have this capability. Thus, in our cases, paradata must be inferred from data entry guides, documentation or communication related to database migrations, technical documentation guiding the structure of the system, and changes to metadata schemas that guide specimen documentation.

3 Database Evolution at the University of Michigan Matthaei Botanical Gardens and Nichols Arboretum and the University of Michigan Herbarium

As part of a larger project studying database maintenance in memory institutions, we developed case studies of database migration at multiple libraries, archives, and museums. Each case study was developed through semi-structured interviews (45–75 min each) with curatorial and collection staff at each site; close analysis and comparison of different versions of legacy databases; and review of papers, memos, emails, and other documentation related to database migration. Evidence is triangulated to develop explanations of how and why migrations are necessary, and to identify patterns motivating migrations, following a multi-case study design (Yin, 2017).

Here we present two intertwined cases from this project: database evolution at the U-M Herbarium and the MBGNA. In both cases, we had access to different versions of the legacy collection databases and were therefore better able to explore different ways of reconstructing paradata. In the process of creating narrative case reports, we also created entity relationship diagrams and Sankey diagrams to better illustrate change over time in these systems. We present our case study and diagrams below.

3.1 Common Origins

The U-M Herbarium is a substantial collection of plants and fungi that began in 1837 (University of Michigan Herbarium, 2023). Today, they have almost 1,750,000 specimens and databasing efforts that span 50 years. Like the herbarium, the MBGNA is essentially a collection of plants—but unlike the herbarium, it is a "living collection" distributed throughout four properties and over 700 acres of land in and around the University of Michigan. The MBGNA's collections and catalogs date back to 1910, and their digital collections databases date back to the 1980s. The MBGNA's digital data collections consist of tens of thousands of items and records in several different database systems. Data files include specimen records describing the type, locality, and provenance of each plant in the gardens and arboretum, as well as images, associated genetic data, and other data files.

Both the Herbarium and the MBGNA databases share a common origin: a holotype specimen database called TAXIR (Estabrook, 1979; Estabrook & Brill, 1969; Estabrook & Rogers, 1966). Holotypes are the single specimen used as a reference when first describing a species (Britannica, 2023). Holotypes might be thought of as the canonical version of a species, and they are incredibly important as a point of reference in biodiversity research. There are many other kinds of "type" specimens that are used in biodiversity research; for instance, paratypes are additional specimens that were referenced when an author described a species, and neotypes are specimens chosen to replace a holotype when lost or destroyed. Collectively, these specimens are referred to as a type collection in a museum. Type collections typically represent a small percentage of an entire natural history museum (NHM) collection (indeed, some NHMs may not have any). However, they are some of the most frequently used specimens and must be extremely well documented to be used as a comparison point for researchers.

TAXIR was a type database created not for the management of these collections but as a research tool to assist in taxonomic classification; it compared holotype records and mathematically calculated whether two specimens are of the same type by analyzing similarities between a pair of specimens (Estabrook & Rogers, 1966). Though TAXIR was an important early NHM database, it simply did not have the capability to serve as a collection management database, or a database where every object within the collection would have its own record, often resembling an analog card catalog. Both the Herbarium and the MBGNA eventually desired collection databases. At this point, both collections consequently diverged infrastructural paths to develop their own databases.

3.2 MBGNA Database Development

After TAXIR, the MBGNA transitioned to the collection management database BG-BASE in the 1980s, a system designed especially for botanic gardens and arboreta collections. This transition was challenging for multiple reasons. Firstly, it was a radical change in the data model as the information documented in a collection database would be vastly different from a type database. Secondly, this migration involved a software and hardware shift, as TAXIR operated on mainframe computer technology, with the data being accessed on computer terminals, and BG-BASE utilized personal computer technology.

Because this transition happened so long ago, our interview participants knew relatively little about BG-Base. However, the MBGNA was still dealing with the ramifications of this software—because they were not able to migrate data out of the system! In our interviews, a curator explained that the data in BG-BASE was originally entered by a volunteer who did not document their data entry process:

> We have a number of databases that people created to, basically to suit their own needs. And so, unfortunately as those staff members have left, we haven't always known exactly how or why those files were created. And so that has been a real challenge to decipher these past records and know what they related to. And sometimes they were created and then updated, and dates weren't kept and so it's really hard to know exactly how to use them (Participant MBGNA-01).

When staff went to migrate the data from BG-BASE to a new system (Microsoft Access) in the early 2000s, the volunteer had moved on, and current staff did not understand the logic behind the data structure. They ultimately had to abandon the BG-BASE database and start fresh. He stated,

> We didn't wanna lose a whole lot of legacy information, but we decided at a certain point it was so erratic that it wasn't worth carrying over . . . it made no sense when you actually just look at the page (Participant MBGNA-02).

The new Access database was created based on the International Transfer Format, a metadata standard for botanical garden data. This standard was developed so that records could be shared easily between institutions (Botanic Gardens Conservation Secretariat, 1987; Botanical Gardens Conservation International, 2004). However, the standard itself is quite complex; the Access database included many small tables with specialized information that link to the main table of the database, aptly named *tblplant*. While this model followed an established standard, it increased the complexity of the database, ultimately making the system harder to use by those without database expertise.

At the time of our interviews with MGBNA staff in 2018, the Access database was being migrated to a new ArcGIS GeoDatabase, a system whose data model prioritizes spatial or geographical information first and foremost. Again, the migration to ArcGIS was unexpectedly challenging because there was little detailed documentation defining fields and relationships in the MS Access database, a similar problem experienced with the BG-BASE database two decades prior. The database specialist for the botanical gardens described this challenge of navigating the table relationships between the two systems:

> When we link Access into [Arc]GIS, the way the data shows up is not the actual characters that were entered into the field originally. It's numbers. Because they're tied to the super relationship table . . . the primary keys show up, that link into GIS (Participant MBGNA-04).

As a result this migration took over 2 years to complete.

Capturing Paradata Through Entity Relationship Diagrams

For the MBGNA case, we had access to versions of the Microsoft Access and ArcGIS databases. To better understand how and what changed over time and to explore different modes of capturing paradata, we developed enhanced entity relationship (EER) diagrams to document each version of the database structures. Creating EER diagrams is a standard practice in database development; they use a highly structured visual language to represent the classes of information in a database, the relationships between those classes, and sometimes the specific data types of different data attributes (Chen, 1976). EERs are commonly created to support a number of different tasks, including database design, debugging, patching, and documentation.

We were able to create two EERs to try to show changes to the MBGNA's structure in the transition from Access to ArcGIS. Figure 2 shows the MBGNA database as it existed in 2008 as an Access database with 20 related tables. Figure 3 shows the database as it was revised in 2016 to an ArcGIS system, condensed into nine tables. Note that the purpose of these figures is not to understand the data structure in detail, but, rather, to note the changes between these two database versions, adding to the paradata on what changes are made over time. There is no

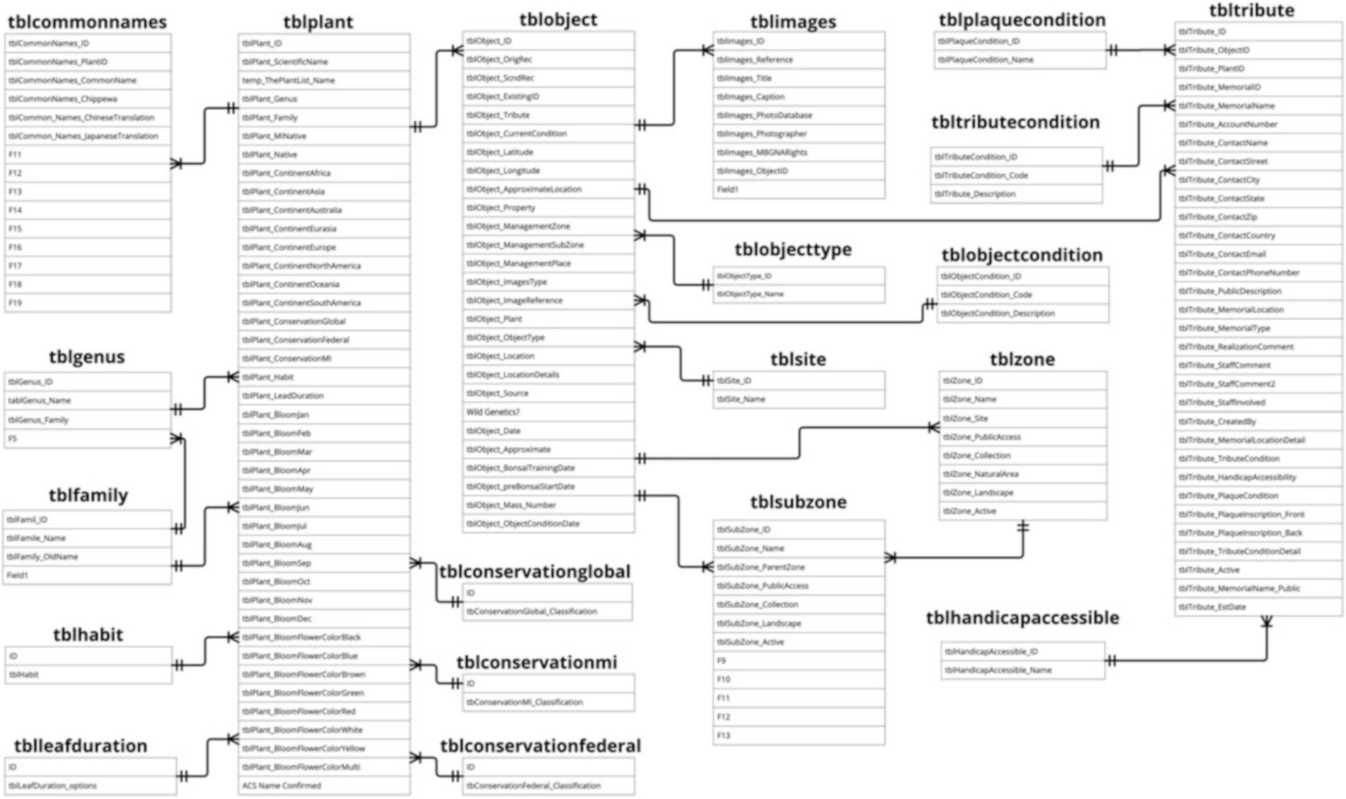


Fig. 2 EER diagram of 2008 MBGNA Access database. To generate the above diagram, we migrated the original Microsoft Access database (.mdb file extension) to MySQL using MySQL Workbench

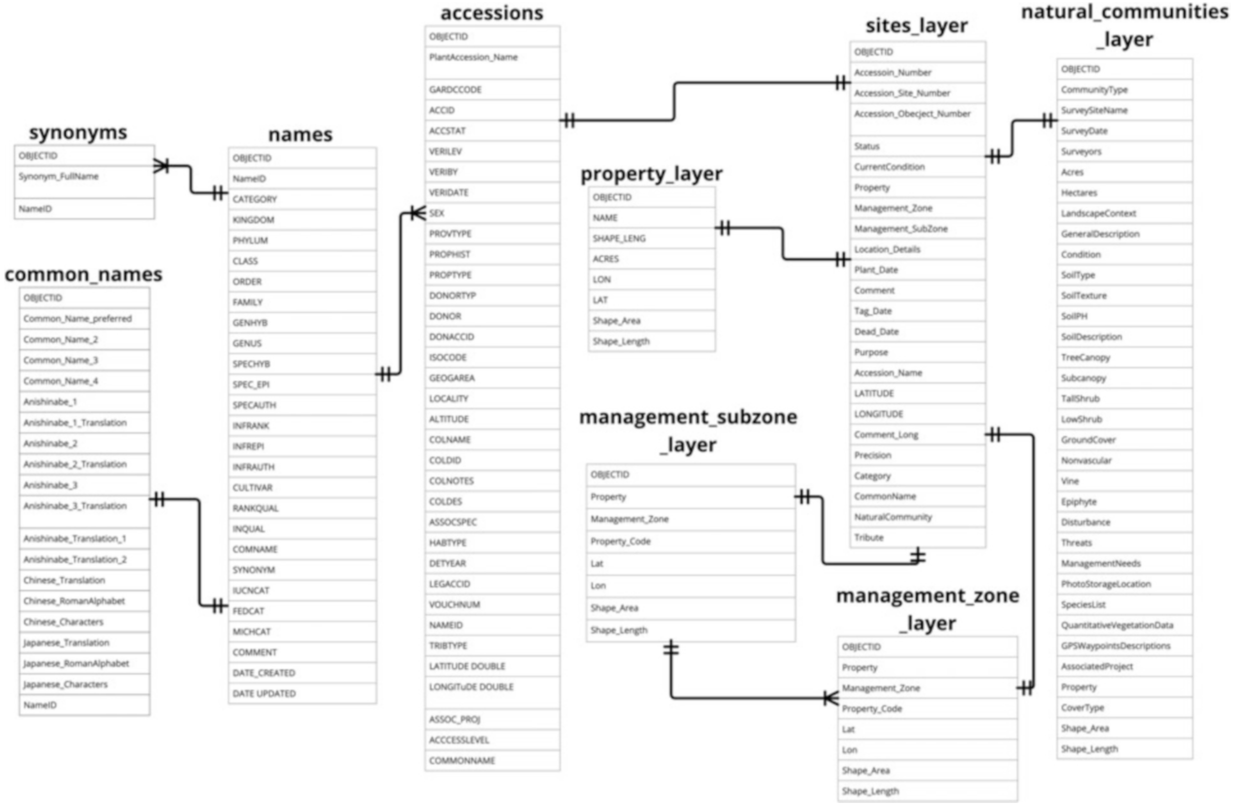


Fig. 3 EER diagram of the MBGNA's current ArcGIS database. We created the diagram by using a Python script to convert the schema details available in JSON into a SQL script, which was then processed using MySQL and MySQL Workbench

EER for the TAXIR or BG-BASE databases because this data was never maintained by staff: a clear example of lost paradata.

A close comparison of these EER diagrams gives a sense of a schema in transition, catalyzed in part by new standards and technologies. The 2008 Access database exhibits design choices tailored to data entry, such as Boolean fields that would have appeared as a checklist to students and interns doing data entry and shows signs of incremental change over time (e.g., several legacy tables that were used for short-term projects, yet were left in after the project was complete, adding unnecessary clutter to the data schema). The most recent database, on the other hand, was built using ArcGIS and represents a large-scale refactoring of the MBGNA's system. It features fewer tables than its predecessors, largely because ArcGIS does not treat controlled vocabularies (or "domains" in their parlance) as separate tables, as the prior databases did. Fewer tables mean simpler queries and quicker data retrieval.

3.3 The U-M Herbarium Database Development

The Herbarium's database evolution differs from the MBGNA's considerably. After TAXIR, their databasing efforts centered around migrating to another type specimen database built in an RDMS called dBase in 1988. This transition was largely due to worries that TAXIR's mainframe technology would become obsolete and out of a desire to access data on personal computers. The Herbarium didn't adopt a collection management database until the early 2000s, after individual researchers began developing ad hoc systems for their own use. One participant describes these systems:

> We had specimen label data... We had some algae data that we came up with from somewhere—we merged that in. I think at one time I had a compilation of 15 or 16 or more databases, that somebody had put a little bit of fungus data here, some algae data here. Stuff that we sort of combined it all together (UMNHM_014).

While decisions were being made regarding what unified RDMS the Herbarium would choose, they stored records in what they called "The Container," which, in the words of UMNHM_013, "ran through SQL with an Access front end, because it was such a monster. It was an enormous flat file intended as a conversion vehicle." But though "The Container" was meant to be a temporary storage solution, the various Herbarium databases wound up spending over a decade in this limbo.

In 2016, the Herbarium migrated to Specify for their in-house collection management database and Symbiota for collaborative research and data aggregation efforts. While the records in these systems are related, one of the challenges that appeared during our interviews is that Symbiota and Specify have historically had issues communicating with one another, largely because of Symbiota's focus on georeferenced data. One participant describes this challenge:

> Symbiota essentially would export out a Darwin Core archive as a big flat table with one exception, and Specify essentially divides up the data into multiple, multiple tables. Getting that all in the right place was very difficult. One of the things that has happened, was as the data is getting out in Symbiota portals, some of the project managers would georeference the data. And you would have georeferenced data out here, [and then realize,] 'Oh I need to get it back in here. Oh, now I have a problem' (UMNHM_014).

In sum, Specify records can be imported to Symbiota, but not vice versa. The provenance challenge here is a lack of documentation to the way in which these two systems (with different intended audiences) fail to interact.

Though the Herbarium database was successfully migrated, the "traces" left behind by legacy databases continue to impact current migration efforts—for instance, fields with unclear definitions, or that were split across multiple tables for unclear reasons. Further, these databases had multiple purposes, including to define type specimens, to serve as a collection database, to be used internally for research projects, and to aggregate collection data with that from other herbaria. This creates a challenge of not only tracing database changes but also intended uses of the systems as well. The Herbarium is fortunate to have a few staff members whose tenure spans since the beginning of databasing efforts and who can help track these changes over time, but this knowledge has not been clearly documented.

Capturing Paradata Through Sankey Diagrams

Because we did not have access to many of the legacy database files for the Herbarium, we turned to alternative methods of visualizing database migrations over time. Sankey diagrams are a type of "flow" diagram commonly used to visualize inputs and outputs in a system, e.g., the flow of nutrients in an ecosystem, or the energy transferred between different components of an engine; in an engineering context, they're used to analyze the efficiency of a given system. Here, we use them to show the flow of records between data systems over time. The basic components of a Sankey diagram are a beginning node, an end node, and a quantitative value linking the two (Schmidt, 2008). For our cases, the beginning and end points for each "flow" are different data systems, and the quantitative value linking the two might be the number of records stored in each system (Fig. 4).

By creating a Sankey diagram, we can succinctly visualize the relationships between older database systems in both the Herbarium and MBGNA, such as the shared use of TAXIR. This diagram also shows the origin of the digital databases in legacy card catalogs (something not possible with an EER diagram). Our Sankey is somewhat incomplete, however, because we don't know the exact numbers of records that flowed through each system. Because of this, we have represented all migrations as being of equal size (and therefore, equal widths of flows linking systems), which does not adequately represent changes in data scale over time. This is an unconventional form of paradata; it might be better thought of as a type of analysis made possible *by* paradata rather than paradata in and of itself. But it does succeed in presenting the provenance of each data system at a high level—something that a new database manager would need when taking over stewardship of a system.

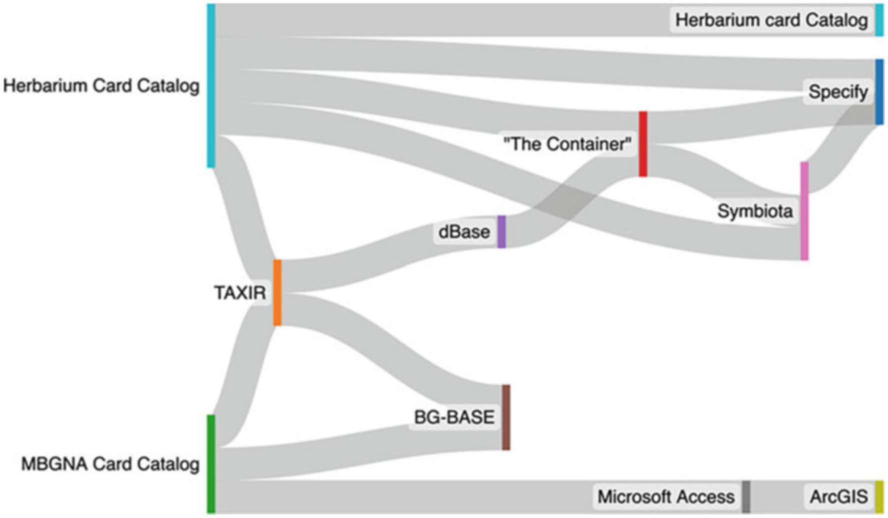


Fig. 4 Sankey diagram showing flow of records between databases at the U-M NHM Herbarium. Diagram created using SankeyMATIC

3.4 Reconstructing Paradata: How and to What Degree Can We Record Change in Complex, Sociotechnical Systems?

Through our interviews, we were able to (mostly) reconstruct the history of these intertwined databases. In doing so, we also created three ways of looking at paradata: a series of EER diagrams that show change when compared; a Sankey diagram that shows the flow of records over time; and the case narratives themselves, which provide a high-level description of the histories of database use. Each of our approaches to capturing paradata records different degrees and scales of change to a system.

EER diagrams are best at capturing detailed changes to the data model, and less directly, the data entry workflow and what fields might be included in the individual record. They are a common tool in database design and management, and provide the clear, detailed views of a system necessary to plan a migration, or understand a current system; however, they can be challenging to create, particularly for those with less experience with databases. And note they are not possible to create for legacy systems that are no longer accessible. Further, these diagrams also require specialized knowledge to interpret. So, while they may be very useful to individuals with backgrounds in computer science, museum administrators or “domain” users may struggle to interpret them. And finally, these diagrams cannot capture why schema design choices were made in the first place.

Sankey diagrams are excellent for providing a broader view of how records have flowed between data systems over time and are even useful for systems that are no longer accessible or that are physical rather than digital. However, if there is no quantitative information on the number of records, or the scale of the data for each system, they run the risk of overly abstracting the complexity of data systems. Also, a Sankey alone lacks much of the behind the scenes work of managing a database, including the context (why and how) of migrations, issues within the databases, and issues with the database schemas.

Finally, we want to note an unexpected form of paradata that emerged from this work: the narrative case reports we developed through interviews and presented in part in this chapter. While these narratives don't necessarily capture the nuanced changes to a data model shown in an EER diagram, or the high-level flow of information between systems in a Sankey, they do preserve much more the social and organizational context of a database that simply cannot be captured through computational methods alone. Databases are sociotechnical objects; as we reviewed before, they are fundamentally shaped by their users and their cultural contexts. Research methods designed to understand people and culture (e.g., qualitative interviews) are needed to surface the human (vs. computational) drivers of change.

We posit that database stewards themselves could create similar narratives, akin to a README. These documents could be a low-barrier way of documenting the changing context and development of a system over time. The strength of a written narrative is its relative ease of creation and comprehension: all database managers have the skills needed to create and read narrative histories of their information systems. Our work here echoes prior work by Bates et al. (2016), Feinberg (2017) Mosconi et al. (2022), and Witt et al. (2009), who similarly use qualitative methods to create sociotechnical narratives of digital systems or objects. By documenting provenance through qualitative narratives, as well as collecting computationally generated paradata, database stewards could create something more like an extended para*text*, akin to a foreword or introduction of a book. All of our participants expressed a desire for documentation that served as a better "threshold," in Genette's (1991) framing, to their inherited systems. Better capture of paradata to show the change in data models and record flows is one part of this, but developing qualitative narratives is likely needed as well.

4 Conclusion: Creating Effective Thresholds to Complex Digital Objects

In this chapter, we have shown several approaches to retrospectively reconstructing paradata and other contextualizing documentation of a complex, long-lived digital object: the museum collection database. At our study sites, the U-M Herbarium and the MBGNA, collection managers face challenges using and migrating legacy database systems because they simply lacked the documentation necessary to understand their predecessors' work. We demonstrated three ways to reconstruct this documentation: through EER diagrams, Sankey diagrams, and narrative case histories developed through qualitative interviews. From a practical standpoint, each of these approaches has strengths and weaknesses but together could be used by database stewards to either document their systems for future generations or reconstruct their databases' histories for their current ongoing management and use. More theoretically, though, we have discussed how paradata, like paratext, is needed as a threshold to a digital object—a way of opening or presenting media to a new "reader." We posit that thinking of paradata as a paratext may open fruitful future research and application avenues.

Acknowledgments Many thanks to our study participants for their time and insights. Thanks to Allison R. B. Tyler for her contributions to early iterations of this work. This project was funded by IMLS grant RE-07-18-0118-18.

References

Bates, J., Lin, Y. W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society, 3*(2), 205395171665450. https://doi.org/10.1177/2053951716654502

Börjesson, I. L., Sköld, O., Friberg, Z., Löwenborg, D., Pálsson, G., & Huvila, I. (2022). Re-purposing excavation database content as paradata: An explorative analysis of paradata identification challenges and opportunities. *Knowledge Creation, Dissemination, and Preservation Studies, 6*(3), 1–18. https://doi.org/10.18357/kula.221

Botanic Gardens Conservation Secretariat. (1987). *The international transfer format (ITF) for botanic garden plant records*. Hunt Institute Botanical Documentation, Carnegie-Mellon University.

Botanical Gardens Conservation International. (2004). International transfer format for botanical garden plant records version 2. https://www.bgci.org/files/Databases/itf2.pdf.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences* (First paperback edition). The MIT Press.

Britannica. (2023). Holotype. Britannica. https://www.britannica.com/science/holotype.

Brodie, M. L., & Stonebaker, M. (2015). *Migrating legacy systems: Gateways, interfaces the incremental approach*. Morgan Kaufmann.

Bruns, T. D., Szaro, T. M., Gardes, M., Cullings, K. W., Pan, J. J., Taylor, D. L., & Li, Y. (1998). A sequence database for the identification of ectomycorrhizal basidiomycetes by phylogenetic analysis. *Molecular Ecology, 7*(3), 257–272. https://doi.org/10.1046/j.1365-294X.1998.00337.x

Buneman, P., Chapman, A., & Cheney, J. (2006). Provenance management in curated databases. In *Proceedings of the 2006 ACM SIGMOD international conference on management of data* (pp. 539–550). ACM Press. https://doi.org/10.1145/1142473.114534

Buneman, P., Müller, H., & Rusbridge, C. (2009). Curating the CIA factbook. *International Journal of Digital Curation, 4*(3), 29–43. https://doi.org/10.2218/ijdc.v4i3.126

Chen, P. (1976). The entity-relationship model-toward a unified view of data. *ACM Transactions on Database Systems (TODS), 1*(1), 9–36. https://doi.org/10.1145/320434.320440

Cullings, K. W., & Vogler, D. R. (1998). A 5.8s nuclear ribosomal RNA gene sequence database: Applications to ecology and evolution. *Molecular Ecology, 7*(7), 919–923. https://doi.org/10.1046/j.1365-294x.1998.00409.x

Dempsey, L. (2000). Scientific, industrial, and cultural heritage: A shared approach: A research framework for digital libraries, museums and archives. *Ariadne,* 22 (22). http://www.ariadne.ac.uk/issue22/dempsey/

Dourish, P. (2017). *The stuff of bits: An essay on the materialities of information*. The MIT Press.

Estabrook, G. (1979). A TAXIR Data Bank of seed plant types at the University of Michigan Herbarium. *Taxon, 28*(1/3), 197–203.

Estabrook, G., & Brill, R. (1969). The theory of the TAXIR accessioner. *Mathematical Biosciences, 5*(3), 327–340. https://doi.org/10.1016/0025-5564(69)90050-9

Estabrook, G., & Rogers, D. (1966). A general method of taxonomic description for a computed similarity measure. *BioScience, 16*(11), 789–793.

Feinberg, M. (2017). Reading databases: Slow information interactions beyond the retrieval paradigm. *Journal of Documentation, 73*(2), 336–356. https://doi.org/10.1108/JD-03-2016-0030

Genette, G. (1991). Introduction to the paratext. *New Literary History, 22*(2), 261–272.

Hine, C. (2006). Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science, 36*(2), 269–298. https://doi.org/10.1177/0306312706054047

Huvila, I., Sköld, O., & Börjesson, L. (2021). Documenting information making in archaeological field reports. *Journal of Documentation, 77*(5), 1107–1127. https://doi.org/10.1108/JD-11-2020-0188

Karr, A. F. (2010). *Metadata and paradata: Information collection and potential initiatives*. National Institute of Statistical Sciences. https://www.niss.org/research/metadata-and-paradata-information-collection-and-potential-initiatives.

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. In *Proceedings of the joint statistical meetings*. American Statistical Association. http://sampieuchair.ec.unipi.it/wp-content/uploads/2018/10/Couper-et-al.pdf.

Manovich, L. (2002). *The language of new media*. MIT Press.

Mineta, K., & Gojobori, T. (2016). Databases of the marine metagenomics. *Gene, 576*(2), 724–728. https://doi.org/10.1016/j.gene.2015.10.035

Mosconi, G., Randall, D., Karasti, H., Aljuneidi, S., Yu, T., Tolmie, P., & Pipek, V. (2022). Designing a data story: A storytelling approach to curation, sharing and data reuse in support of ethnographically-driven research. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–23. https://doi.org/10.1145/3555180

Palmer, C., Weber, N. M., Renear, A., & Muñoz, T. (2013). Foundations of data curation: The pedagogy and practice of "purposeful work" with research data. *Archives Journal*. Retrieved November 20, 2019, from http://www.archivejournal.net/essays/foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-data/

Pomerantz, J. (2015). *Metadata* (The MIT Press essential knowledge series). MIT Press.

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., & Desmet, P. (2014). The GBIF Integrated Publishing Toolkit: Facilitating the efficient publishing of biodiversity data on the internet. *PLoS One, 9*(8). https://doi.org/10.1371/journal.pone.0102623

Schmidt, M. (2008). The Sankey diagram in energy and material flow management. *Journal of Industrial Ecology*, 12(1), 82–94. Retrieved December 15, 2019, from http://onlinelibrary.wiley.com/doi/10.1111/j.1530-9290.2008.00004.x. https://doi.org/10.1111/j.1530-9290.2008.00004.x

Stein, B. R., & Wieczorek, J. R. (2004). Mammals of the world: MaNIS as an example of data integration in a distributed network environment. *Biodiversity Informatics, 1*, 14–22. https://doi.org/10.17161/bi.v1i0.7

Strasser, B. J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(1), 85–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22326076. https://doi.org/10.1016/j.shpsc.2011.10.009

Thomer, A. K., & Rayburn, A. J. (2023). 'A patchwork of data systems': Quilting as an analytic lens and stabilizing practice for knowledge infrastructures. *Science, Technology, & Human Values*. https://doi.org/10.1177/01622439231175535

Thomer, A. K., Weber, N. M., & Twidale, M. B. (2018). Supporting the long-term curation and migration of Natural History Museum collections databases. *Proceedings of the Association for Information Science and Technology, 55*(1), 504–513. https://doi.org/10.1002/pra2.2018.14505501055

Thomer, A. K., & Wickett, K. M. (2020). Relational data paradigms: What do we learn by taking the materiality of databases seriously? *Big Data & Society, 7*(1), 205395172093483. https://doi.org/10.1177/2053951720934838

University of Michigan Herbarium. (2023). About us University of Michigan Herbarium. About University of Michigan Herbarium. https://lsa.umich.edu/herbarium/about.html#:~:text=The%20Herbarium%27s%20collections%20were%20initiated,the%20world%2C%20to%20its%20care.

Vieglais, D., Wiley, E. O., Robins, R., & Peterson, T. (2000). Harnessing museum resources for the Census of Marine Life: The FISHNET Project. *Oceanography, 13*(3), 10–13. https://doi.org/10.5670/oceanog.2000.02

West, B. T., & Sinibaldi, J. (2013). The quality of paradata: A literature review. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 339–359). Wiley. https://doi.org/10.1002/9781118596869.ch14

Williams, J. W., Grimm, E. C., Blois, J. L., Charles, D. F., Davis, E. B., Goring, S. J., Graham, R. W., Smith, A. J., Anderson, M., Arroyo-Cabrales, J., Ashworth, A. C., Betancourt, J. L., Bills, B. W., Booth, R. K., Buckland, P. I., Curry, B. B., Giesecke, T., Jackson, S. T., Latorre, C., . . . Takahara, H. (2018). The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quaternary Research*, 89, 159–177. https://doi.org/10.1017/qua.2017.105

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation, 4*(3), 93–103. https://doi.org/10.2218/ijdc.v4i3.117

Yin, R. K. (2017). *Case study research and applications: Design and methods*. Sage.

Alexandria J. Rayburn is a doctoral candidate at the University of Michigan School of Information, where she also received her Master's in Information Science. Her research examines computing practices in museums and knowledge infrastructures. In particular, Alexandria is interested in the maintenance of digital systems and the labor practices associated with that maintenance. Her research has been published in Science, Technology and Human Values, Archival Science, and more.

Andrea K. Thomer is Assistant Professor at the University of Arizona School of Information. Her research interests include the maintenance and evolution of knowledge infrastructures, scientific data curation, and information organization. She is especially interested in long-term database curation, the use and impact of natural history collections, and the conceptual foundations of data science. Dr. Thomer earned her doctorate at the School of Information Sciences at the University of Illinois at Urbana-Champaign in 2017. Prior to her graduate work, she was an excavator and ad hoc data curator at the La Brea Tar Pits in Los Angeles, California.

Paradata in Emergency Services Communications Systems[*][†]

Megan Cohen, Jardi Martinez Jordan, M. Scott Sotebeer, and Michael Stiber

Abstract

Government provides a variety of critical services, often grounded in decades-old technology. As these services evolve to encompass newer technologies and offer broader capabilities, their exposure to potential threats increases the importance of modeling and simulation to understand their operation and vulnerabilities. This modeling activity requires access to records collected by government agencies, conceptualized formally in terms of archival science. However, the archival science concept of paradata is insufficient to support the forensic reconstruction of the real world required by a modeling effort. This chapter outlines how iterative, interdisciplinary work to model and simulate emergency services communication systems illuminates the need for a new role for paradata in government archives.

[*] This work was funded in part by the National Security Agency under Grant Number H98230-20-1-0314. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation herein.

[†] This work is in part an outcome of InterPARES Trust AI, an international research partnership led by Drs. Luciana Duranti and Muhammad Abdul-Mageed, University of British Columbia, and funded by the Social Sciences and Humanities Research Council of Canada (SSHRC).

M. Cohen · J. M. Jordan · M. Stiber (✉)
Computing & Software Systems Division, School of STEM, University of Washington, Bothell, WA, USA
e-mail: jardiamj@uw.edu; stiber@uw.edu

M. S. Sotebeer
USA Strategics, Jacksonville Beach, FL, USA
e-mail: scott@usastrategics.com

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_10

1 Introduction

When a critical incident happens—a fire, medical event, natural disaster, crime in progress—we expect to be able to take out our phone, dial a short number, and almost immediately be paired with a person who will quickly send us the help we need. In fact, there is increasing demand for a broader range of emergency services communications systems (ESCS), broadly named “next generation systems.” The public expects these tools to be fully current and compatible with available technology. Call receivers and dispatchers are now expected to not only answer wireless and hardline calls, but to also intake video, photos, and text. To evaluate these systems, however, it is critically important to first understand what exists at the present, how it functions, how it can be optimized and improved upon, and where its most critical vulnerabilities and points of failure (from cybersecurity threats, natural disasters, technical problems, etc.) are.

Our research, Emergency Communications and Critical Infrastructure, aims to both understand and optimize existing emergency call and response processes by coupling large-scale simulations, existing, real-world datasets generated from government records management systems and practices, and artificial-intelligence-driven data analytics tools. Through this multiyear effort, we are developing generalized abstract models of ESCS, focusing initially on the “911” system in North America (and more specifically, the United States), building a computer environment capable of simulating ESCS at large scale (city to national level), validating these simulations against real-world data, using the software created for training and “what if” scenario exploration, and developing tools that examine call data in real time to perform tasks such as prioritizing calls when systems are overloaded and monitoring community health changes. Aspects of this work are supported by government (critical infrastructure) and by information science collaborators (data provenance and applications of AI in records and archives).

Like most, if not all, government functions, emergency communication management has workflow and documentation built into its process. Records Management Systems are a necessary and inherent byproduct of the activity, legally required for evidentiary and other public records purposes. This can mean that records may act as evidence of crimes committed, but also as evidence that the activity (in this case, emergency phone calls) itself is happening, or that the agency is functioning at all. These records can be compiled into large datasets that form the basis of our research.

These two activities—academic research and ESCS operations— each generate their own sets of paradata. Thus, this chapter is concerned not only with these two forms of paradata, but also how both of them inform the modeling and analysis efforts.

While ESCS in many North American jurisdictions look similar, they are not identical. For the purpose of the first phase of our research, we particularly focused only on priority-based, critical incident, nonbusiness calls (911 calls from citizens) in King County, Washington, USA, which was selected via processes discussed later in this chapter.

2 Paradata in Archival Science

The term "paradata" first appeared in the social sciences as a way to describe data collected during surveys and interviews that were, in some way, extra (Kreuter, 2013). This data, often described as marginalia or annotations, is generally created and used by researchers in the process of first collecting and later analyzing their data and have long been recognized as critical elements of research in both qualitative interviews and survey methodologies (O'Connor & Goodwin, 2017; Schenk & Reuß, 2023). Because of the way it is created, paradata is often referred to simply as "process data."

As computers became a common tool for conducting surveys, the kinds of information that could be considered paradata changed. Common examples of these new kinds include mouse movement information and keystroke response time data, which can only be collected by a computer and which have been found to be so meaningful that they have warranted research in and of themselves (Fahmy & Bell, 2017). Paradata has even been used as the basis for machine-learning-based predictive analyses (Fernández-Fontelo et al., 2020). While paradata is an accepted phenomenon in some humanities and social sciences, it has not yet taken hold in archival science, where the distinctions among data, metadata, and paradata remain unclear.

In part, this is because many of the original examples of paradata, such as annotations and marginalia, have long been recognized as elements of records in the study of archival diplomatics (as discussed also in Trace and Hodges, 2023). Archival theory has a great deal to say about the role of annotations in the execution, handling, and management of records (Duranti, 1991). But while annotations have traditionally been considered an extrinsic element of records, defined, and examined by form rather than content, the term "paradata" instead considers annotations by their intrinsic and qualitative value, by content rather than form. We can see this in the InterPARES Project's definition of paradata: "information about procedures and tools used to create and process information resources, along with information about the persons carrying out those procedures" (Davet et al., 2022). Archival science traditionally has very little to say about the content of a given annotation or information resource, and this departure is significant.

Paradata, inherently, is less relevant to the organization of an information resource than it is to the interpretation and reuse of that resource, either by its creator or by a third party at a later date. Most examples of paradata (field notes, annotations, mouse click data, etc.) are in the form of supplementary data within an initial collection process—marginalia, for example, must inherently be in the margins of another, likely more official or formal, document. Paradata, if defined as any information relevant to the creation processes or persons, is a far broader category of information than early definitions of "extra" or "exhaust" data (Pomerantz, 2015). It is entirely reasonable that information intentionally collected in the regular course of government activities could inform the acts and persons of creation. (For example, the identity of the call technician in an emergency call is intentionally recorded and stored and is paradata.) Conceiving paradata as something exclusively extra or unintended, therefore, is a mistake.

Now the lines between data, metadata, and paradata are blurry to the point of being indistinctive. That is not a bug but a feature—if data and metadata are defined by their process of creation, and paradata is defined by its informational content, there will inherently be some overlap.

Some data is paradata, and some metadata is paradata.

3 Paradata in Modeling the Real World

The researchers here are not data creators, at least in terms of the actual ESCS operation ("real world data"). Instead, researchers were granted access to a (modified) pre-existing dataset to manipulate and interpret. We needed paradata, either what was interpretable from the metadata or what could be learned through interviews and ongoing contact with government agencies, to inform the modeling process.

Modeling and simulation not only seek to understand the world but also construct a formally defined and precise representation of the world and implement that representation in software. Thus, there is a sequence of complex links in the chain from the real world to the software and, moreover, that final step to software brooks no ambiguity and implements exactly any discrepancies that have crept into the research process.

As a result, this chapter focuses on two separate forms of paradata within this research. First is paradata generated by researchers over the course of the research itself. Examples include documentation of the iterative development of the modeling workflow and meeting minutes among researchers. The second is paradata that the researchers seek out, created by others via processes beyond the researchers' control, to understand the context of creation of the records and dataset that they are modeling.

The goal of the models in this research is to, as much as possible, recreate the real-world conditions and transactions that generated the dataset. The data itself is not sufficient to accomplish this; we need to understand the processes, policies, decision-making procedures, and people involved in the process. This information might be called "forensic reconstruction paradata."

3.1 The Map Is Not the Territory: Forensic Reconstruction of the Real World

Over 240 million, 911 calls are made in the USA each year. The private sector drives the business of 911 in terms of technology and hardware and software infrastructure. Federal, state, and local governments drive the day-to-day implementation and operation of the system. While trade and quasi-regulatory organizations such as the US National Emergency Number Association (NENA) capture, coordinate, and catalog the national operation and policies around 911 generally, there is no centralized and coordinated US clearinghouse for baseline hardware and software standards, data management, R&D, or privacy standards, among other 911 policy issues.

Unlike the European Union and its harmonized regulations, or Israel and its single, national emergency call receiving and dispatch system, the USA and its over 6,100 call centers represent a patchwork of independent policies specific to their unique unit-of-government structure. The cascading effect related to 911 call data alone relates directly to issues such as states' rights down to various regional (counties) and local government (municipal) regulatory structures (cities, townships, fire districts, etc.) and the philosophical, political, and mostly emotional concept of "local control."

Thus, the paradata collected in the process of operating 911 primarily serves 911 call centers' needs. Moreover, regardless of current technology, 911 call and dispatch is first and foremost reliant on human behavior. Call takers' actions, responses, and interactions, both formal and informal, are grounded in their training. Paradata within the call intake and dispatch process can be as simple as the screen notes made by a single individual call taker who is also dispatching. These "notes" may or may not become part of a call record—either due to technical limitations or local data retention policies.

This paradata may also be incredibly complex. A single event may involve many call takers and dispatchers in the same public safety answering point, or PSAP, who may communicate informally by voice or email. While these notes and interactions among PSAP members may have a direct influence on decision-making in real time, they are not likely to be captured as paradata for research purposes.

However, it is certainly the case that the value of paradata in emergency response is recognized as critical to "after action reporting" and to day-to-day call center management. Given that call centers operate 24/7 and top management is not present for two of three shifts, notes and recollections of supervisors and call receivers and dispatchers directly influence personnel management as well as operations and call process, evaluation, and policy.

Therefore, while the formal records can still suit the needs of research, these records are not enough to create the forensic reconstruction that will actually address the research questions. It is more likely that activities such as interview "deep dives" will be needed to understand relevance to the research process (Simpson, 2020).

3.2 Iterative Workflow

Since this research does not require researchers to generate their own real-world datasets but rather collect existing ones, the workflow and process for obtaining data have evolved over time. This is integral to the process of constructing an abstract model and then realizing that model as simulation algorithms and software in some programming language. Each stage forces the researcher to make explicit what might have been implicit and reveals gaps in understanding or a collection of real-world data, metadata, or paradata. This is one of the primary benefits of modeling and simulation: It admits no fuzziness of thought. The current workflow has evolved and may well evolve going forward as we interact with more government agencies, seek out different datasets, and have new gaps in our understanding revealed by the modeling and simulation process. Our workflow, like our understanding of the datasets we work with and the needs of our simulations, develops iteratively.

This also means that while the workflow below is presented in a specific order, that order has not always been the case. For example, some interviews of emergency management officials were conducted before any jurisdiction was selected to request data from, simply so that researchers could begin to understand processes surrounding emergency communications record creation.

Workflow

1. Identify geographic region of interest:
   (a) Determine characteristics of desirable geographic area (for example, urban, suburban, rural).
   (b) Determine critical metadata elements for simulation.
2. Determine legal restrictions on collecting 911 data, including federal, state, and local frameworks:
   (a) Research federal, state/provincial, and local regulations.
3. Determine formal processes for collecting data, and determine the difference between the formal process and the "way it is done":
   (a) Conduct outreach and interviews with employees and managers to understand the processes of creation and capture of records, of creating and storing large datasets, and of access by the public.
   (b) Select storage methods and access restrictions according to legal and ethical standards.
4. Obtain and begin work with a given dataset.

Each of the steps along this workflow produces, in turn, paradata of its own in various forms. For example, the process of selecting a given region for data collection is a group decision and takes place during a group meeting, where formal minutes are taken by the research lead. Those minutes are sent to each member of the team for future reference. In addition, individuals may take their own notes, which may or may not be shared with other researchers. In some cases, decisions in this step have been formalized with documentation that shows a more academic, or literature-justified, intention behind the decision. Step two is also documented. The researcher who finds the information puts the sources and a summary of the research into a document or perhaps presents the information at a group meeting.

Step three, the interview process, is where recordkeeping and documentation become more complicated and nuanced. Interviews necessarily generate a lot of documentation. There may be emails, for example, to set up the meeting time.

Researchers come to interviews with prepared questions that they have each put together, and they may send these in emails to the interviewee ahead of time. Researchers take their own notes, in addition to the formal write up created by the head researcher. There may also be emails sent after the interview to forward resources, send thanks, or follow up for clarification purposes.

A final step in this process that is not outlined in the workflow but is critical to the success of the research is the evaluation of the data that comes once it has been worked with (for example, to create, refine, or expand a model or simulation design). What we learn as a group is based on how usable the dataset is, whether or not we got the metadata we needed to create an interpretable and implementable model, or if there were any unforeseen lessons from working with a given set. These meetings and conversations can take place over weeks and happen between different clusters of researchers, which means that some kind of documentation is critical to the iterative improvement of data collection processes (in addition to the extensive documentation that is generated by any modeling or software development processes).

3.3 Obtaining King County 911 Call Data

Early in this research, King County was selected as a region that could be helpful to work with. It was selected for a few reasons beyond simple convenience. King County is made up of multiple municipalities and various independent units of government (for example, some PSAPs include fire districts with separately elected boards and officials). It includes a major city (Seattle, Washington) as well as less densely populated areas. It is a region with a potential for high variability of data. There are twelve separate 911 PSAPs, including university, fire district, fire department, police departments, police and fire cooperatives, and a county Sheriff's office. There are unique and distinct units of governance involved (such as municipalities, universities, fire districts, Sheriff, etc.) that in turn represent separate and independent call data collection policies, methods, and databases.

This decision was made in the autumn of 2020 but was not recorded in those terms until almost a year later, when additional researchers had joined and were searching for other jurisdictions to begin soliciting data from. The justification for choosing King county was ultimately recorded in one researcher's personal, paper notes.

The identification of preferred metadata elements took place over a number of meetings between different combinations of group members and was recorded entirely on personal devices or as personal notes, likely by multiple members of the team.

Research into the legal requirements and regulations surrounding emergency communications data created a few pieces of paradata. Some of it came from notes taken during interviews with county officials. Some came from formal research and documentation undertaken by individual researchers.

Through the aforementioned interviews, researchers learned how to get access to a King County dataset and did so. Record of this process was also kept in meeting minutes, or in personal notes. The logic behind decisions on where the King County dataset was to be kept, and to whom access would be granted, was not recorded.

Finally, paradata that refers to the progress being made by working with the dataset shows up in a large number of places: in meeting minutes, in private notes, in plenary reports, where researchers present regularly about the status of their findings, and even in conference posters. Lessons learned through this process can also be found in the secondary and tertiary iterations—for example, the lists of desired data, metadata, and paradata have changed as researchers have had more working time with datasets and developed more refined models and software.

4 Paradata and Interdisciplinarity

A pattern noted above is a certain amount of chaos in the paradata from this research. No small amount of this comes from the number of different researchers and their differing areas of expertise. This research is inherently interdisciplinary, bringing together experts in the fields of computer science, machine learning, archival science, cybersecurity, emergency systems management, and critical infrastructure management. These fields are diverse and bring diverse concepts and taxonomies into an already complex effort.

Initial meetings with the full research team (including those directly associated with the InterPARES Trust) were overcomplicated by linguistic confusion between group members. For example, archival science, as noted earlier in this chapter, has a very specific definition for what counts as a record, and how a record can be differentiated from a datum, or a dataset. Language between archivists and computer scientists needed to be discussed at length and over multiple meetings for researchers to be able to communicate effectively. Similar discussions also occurred between the computer-focused researchers and those with experience in critical incident response communications management. These problems are not specific to paradata, but an inevitable element of interdisciplinary research that spilled into the paradata: Not all of the notes, marginalia, minutes, etc., are necessarily written in a manner that is intuitive to every group member.

What is more is that, since this is a long-term effort, spanning multiple years, not every researcher was involved when it began, nor will every researcher still be working on it when it ends. When this work began, before InterPARES was involved, the scope of work and expertise involved was much smaller. The duration of this work and the potential rotation of researchers, research assistants, and other participants will surely complicate the documentation process across the waves of anticipated data gathering and collaborations.

Another critical element of this diversity is a diversity of physical locations. Researchers meet mostly online, communicate online, and keep research materials in online data storage. Data and paradata are stored in mutually accessible digital spaces for all to access. That does, however, inherently limit what paradata can look like. For example, while Google Docs does have a "comment" function on documents, those comments are harder to write than typical marginalia might be and can be later deleted by any party with editing authority, meaning some amount of this paradata is inevitably lost to efforts to keep documents "clean." In the cases where individuals keep their own notes in a freehand manner, this may not be such an issue, but the geographical disparity between researchers means that those notes are not likely to be of help to any other members of the team.

As the data gathering in this research is iterative, our paradata is being actively reused as part of ongoing data retrieval efforts. We are not done yet, and this means that our paradata is particularly useful to us as we get deeper into the research. It also means that we cannot yet grasp the totality of the role our paradata will play in our results.

5 Paradata in Developing Simulations

In fiction, a character like Hari Seldon (in Asimov's *Foundation* series) may be able to accurately predict the fall of a galactic empire using an algorithmic science of his own making. In real life, we cannot yet predict the fall of a civilization, but we do have models that allow us to forecast the weather with some accuracy. Most people are familiar with this predictive nature of modeling. But what is a model? And, how does a computer scientist, foreigner to the inner workings of the 911 system, develop a mathematical model for it?

A pertinent definition of a model is that offered by Pidd (1999) within the context of Operations Research and Management Science: "A model is an external and explicit representation of part of reality as seen by the people who wish to use that model to understand, to change, to manage, and to control that part of reality in some way or other." This definition grants us some important characteristics; a model:

- Can be examined, challenged, and be formally defined
- Is a partial and simplified representation of reality
- Is dependent on the viewpoint of its stakeholders, and
- Is fitted for a specific purpose, therefore goal-oriented

Given those characteristics, it is not surprising that there is no prescriptive methodology for modeling, a process described by Holland (2000) as the act of extracting regularities from incidental and irrelevant details. In the context of this research, our first instinct would be to dive into a 911 call log looking for regularities but without an understanding of the processes that generated it; as a result, our prospective insights will be limited. For instance, we can determine the pattern of call arrivals, service time, and wait times from a call log, but we would not be able to infer the staffing policy of a given PSAP. Unlike machine learning (ML) models that are black boxes where prediction is the primary interest, simulation models are meant for emulating the behavior of the system being modeled to elucidate greater understanding of that system.

To illustrate this modeling process, let us take a look at the six principles of modeling suggested by Pidd (1999):

1. Model simple; think complicated.
2. Be parsimonious; start small and add.
3. Divide and conquer; avoid megamodels.
4. Use metaphors, analogies, and similarities.
5. Do not fall in love with data.
6. Model building may feel like muddling through.

Not surprisingly reminiscent of the empirical model of science, these six principles outline an iterative process where abstraction, simplicity, and decomposition are key aspects; careful thought and analysis drive the collection of data and not the other way around. As we learn more about the 911 system, we are able to formulate better questions and determine what datasets we need. At the time of writing, the relevant dataset of interest includes caller data, responder data, and Geographic Information System (GIS) datasets. The identification of these arose from conversations with stakeholders in addition to technical documents that outline the policies and procedures of the 911 system. For instance, the decision to look into the GIS datasets was due to the call routing and dispatching procedures. Calls are routed based on their geographic location and the service boundaries of PSAPs; moreover, dispatching is dependent on the first responder proximity to the emergency event.

5.1 Modeling Emergency Services Communication Systems

Human activity systems, such as ESCS, have the following characteristics: boundaries, components, behavior, an internal organization, human activity, human intent, openness to the environment, limited life, and self-regulation (Pidd, 2007). Following the previously discussed modeling principles, we address these characteristics in our models through simplification and abstraction. We begin by identifying the main components of the system and then move into modeling three main characteristics:

1. The internal organization of the system
2. The internal behavior of its main components, and
3. The interactions among these components

Although ESCS are complex multinetwork systems, we must shear away details because a model must be simpler than the real system (to be useful). The seemingly simple act of dialing 911 triggers a sequence of steps that involve many layers of technology and emergency personnel, generating data and governed by paradata at each layer. However, we have identified three main types of entities: Caller Regions (CR), Public Safety Answering Points (PSAPs), and Responders. In our models, we use the name Caller Region (CR) to denote a geographic area where calls originate from, a PSAP is an emergency call center responsible for answering emergency calls and dispatching first responders, and Responders indicate the headquarters where first responders are dispatched (e.g., police and fire stations). Moreover, these components are arranged in a network that underlies the GIS-based call routing and dispatching dynamics.

To model ESCS networks, we leverage mathematical constructs known as *graphs*. Graphs are used to model network nodes (called *vertices*) and their connections (called *edges*) in a pairwise relationship. This particular model is a *directed graph* where vertices denote the aforementioned ESCS entities and edges represent the communication channels between them. In a directed graph, the relationship expressed by the edges has a direction (for example, every call has a caller and an answerer, and thus the communication is asymmetrical). Because the connectivity is based on geographic coordinates and boundaries of some of the components, we extract the network topology from GIS datasets by encoding the components' jurisdictional and neighboring relationships in a directed graph.

The abstract concept for modeling the internal behavior of ESCS entities comes from the realization that PSAPs are specialized call centers. Discrete Event Queuing Models, extensively used in the management of call centers, are suitable for modeling the processing of emergency calls by PSAPs. Furthermore, the same concept can be applied on the responders' side for modeling dispatching and response actions (this will not be discussed in this writing as it is not meant to be a complete model description).

To illustrate the concept, we will use the general queuing representation of a call center, shown in Fig. 1. Conceptually, a call center contains k trunk lines with up to the same number of workstations ($w \leq k$) and agents ($n \leq k$). One of three scenarios occurs when a call arrives: It is answered right away if there is an available agent, the call is placed in a queue if there are no available agents, or the caller receives a busy signal if there are no trunks available. In this model, we think of an agent as a resource that is occupied, while a call is being answered, then immediately released once it has been served. Calls are lost due to blocking when all trunks are busy or

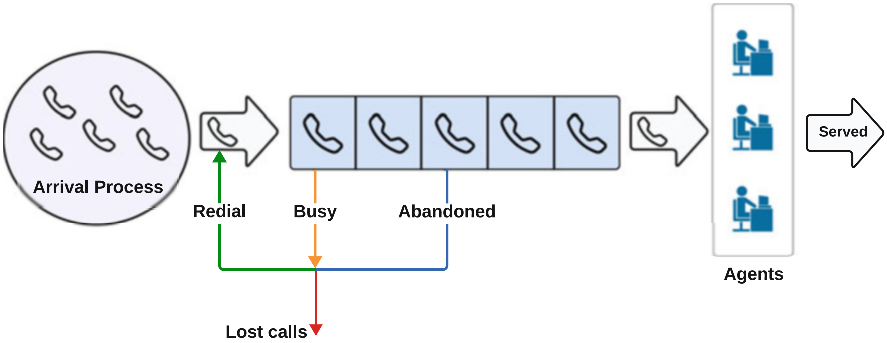


Fig. 1 Call center as a queuing system. The diagram represents a system with 3 agents (n) and 8 trunk lines (k); therefore, the size of the waiting queue is 5 ($n - k$)

when a caller abandons the queue due to impatience, possibly redialing soon after. Consequently, arriving calls come from either those who make an initial call, those whose got a busy signal, or those who abandoned the queue after waiting.

Stochastic processes such as call arrivals, service time, and customer impatience must be modeled through random variables drawn from a suitable probability distribution. Existing research in this area provides possible candidates, but their goodness of fit has to be evaluated based on data obtained from the real system. Call arrivals, in particular, exhibit burstiness—intraday, interday, and seasonal variability—that requires extra modeling effort. One idea for modeling call arrivals comes from the realization that calls are the consequence of emergency events; following this logic, one can model the arrival as a *cluster point process* (Cox & Isham, 1980) characterized by:

- A *primary process* that defines the emergency events as the realization of a stochastic process
- A *subsidiary process* that defines the number of calls triggered by each emergency event and the separation among them through discrete probability densities; and
- A *pooling* that consists of the superposition of all clusters that results in the cluster point process

The last abstract concept in the model is that of Communicating Finite State Machines (CFSM). A simulation starts with an initial setup with defined parameters and runs for a defined number of time steps; the values of parameters and simulation variables at a given time step are known as the *state* of the system. In the model being discussed, the state of the system is the compound state of every vertex and edge in the graph. At every time step, ESCS entities consume inputs and undergo state changes that might cause them to send outputs to other vertices.

For performance purposes, our in-house simulator (Stiber et al., 2017; O'Keefe et al., 2022) is designed to facilitate the implementation of simulation models on Graphic Processing Units (GPUs). These GPU implementations allow us to achieve high-performance simulations of large complex systems, but at the same time, their high parallelization presents challenges for interconnections. Modeling the interaction between ESCS entities as Finite State Machines that communicate with each other by transferring event messages through their connecting edges provides a useful abstraction for highly parallelized processes such as our GPU implementations.

5.2 Iterative Paradata

Thus, we see the close of the loop in the iterative process described herein. What at first seems straightforward—ESCS involves emergency calls—becomes more involved as specific decisions are required in the model and its implementation. Roughly speaking, the road traveled has been:

Paradata Iteration Workflow (Reconstructed)

1. Start with a basic understanding of the system and initial questions/goals for inquiry.
2. Consult with stakeholders, refine conceptual model of system operation, and secure sample call data.
3. This new information reveals new questions for the simulation to answer and that the data cannot be treated from a simple "black box" point of view:
   (a) Stakeholders are interested in inferring the state of the world in real time as calls come in, and so a simple statistical model of calls as a sequence of random events is insufficient. Instead, the model of call generation should include the "primary process" of events, each of which can generate one *or more* calls.
   (b) Call data includes substantial information, such as locations, unique call identifiers (substituted for phone numbers, to identify repeat calls), call type, etc. The metadata surrounding such information is not interpretable without understanding the ESCS processes that govern such things: how caller location is determined, how calls are categorized by PSAPs (free form text, selection from a limited set of choices, decision tree), etc.
   (c) Original call data may include personally identifiable information. Local laws and procedures (which may vary from one jurisdiction to another) control what parts, if any, are public information and the processes for data sharing (Coyle & Whitenack, 2019). The investigators must develop a data management plan that applies sufficient control to satisfy external stakeholders so that memoranda of agreement for data sharing can be created and agreed to by both organizations:
4. The model is expanded to include first responder dispatch, and conversations with stakeholders reveal details of how this dispatch occurs and what data is collected as part of that process.
5. The responder dispatch process is based on geographic locations of entities (callers and responders) and PSAP service boundaries. Therefore, GIS data must also be incorporated into the simulation, and agreements must be reached with external stakeholders to share such data. Luckily, there are widely adopted standards for GIS.

6 Conclusion

The traditional archival understanding of the contexts of records or collections is composed of two concepts: *respect des fonds*, the idea that every set of records can be traced to a single creator or creating body, and original order, the idea that the order that the creator put records into is meaningful and should be preserved. By preserving the identity and order of the creator, the context and meaning of records can be preserved. This forms the archival concept of provenance, a tool that facilitates the arrangement and description of records for archival preservation and access. Provenance focuses, again, not on the content or information within materials but on their contexts of creation and storage by creators, and perhaps for electronic records static structure or metadata.

This traditional approach may be most useful for archivists who manage records like those of government agencies, which tend to fit neatly into static structures and corporate ideas of creatorship. Forms of records and information, however, have changed and continue to change in the digital age. Records and information are so often decontextualized into data, where they can be easily analyzed but not so easily understood. Analysis can combine and recombine data in a vast number of different ways, including AI algorithms that render provenance opaque (as is also discussed in Trace & Hodges 2023). But what is the purpose of analysis if not to create understanding? And shouldn't the construction of models of systems' operation, implementation of such models in software, and analysis of the results of simulation be considered core to our process of understanding?

Records are representative of the thing they document. In the case of the work described in this chapter, everything we are modeling is itself a process—making a phone call happens over time, emergency response vehicles take time to get from A to B. The model we have created, therefore, is itself an embodiment of paradata, process data. Such models can allow researchers to ask questions not possible to ask of the original dataset, and there is unique utility in using models to recreate processes as opposed to querying a database or spreadsheet.

While paradata can help record creators, it is especially important for users, especially if those users are not the original creators. Our research group is a good example here, with diverse backgrounds and needs that caused us to focus on different elements of ESCS systems. The archival understanding of context, creation, and provenance needs to expand.

Paradata in archival science could come to add a third dimension to provenance, as archivists come to realize that structural metadata, the name of the creating agency, and the order in which records are stored may not be enough context to make records usable, either to actors within a given agency or, as in this case, to researchers.

There is limited utility in litigating whether a given piece of information is data, metadata, or paradata. These terms, as previously discussed, are not mutually exclusive. This project shows how this can be a feature. A flexible mindset can allow users of information resources to adapt those resources to answer their specific questions. In our case, where our initial questions concern real-life processes rather than outcomes, paradata is essential. The records themselves may not answer the questions that a model can. This can help drive change, in this case framing a more structured process in ESCS to evaluate, adjust, improve, advance, and ultimately uptrain emergency call receivers and first responders. *Respect des fonds*, original order, and paradata together create a much more complete view of the process of *how* a record or dataset came to be than the first two alone.

References

Cox, D., & Isham, V. (1980). *Point processes*. Chapman and Hall

Coyle, E. K., & Whitenack, S. L. (2019). Access to 911 recordings: Balancing privacy interests and the public's right to know about deaths. *Communication Law and Policy, 24*(3), 307–345. https://doi.org/10.1080/10811680.2019.1627796

Davet, J. E., Hamidzadeh, B., Franks, P. C., & Bunn, J. (2022). Tracking the functions of AI as paradata & pursuing archival accountability. In *Archiving Conference, Society for Imaging Science and Technology* (pp. 83–88). https://doi.org/10.2352/issn.2168-3204.2022.19.1.17

Duranti, L. (1991). *Diplomatics: New Uses for an Old Science, Part V*. Archivaria https://archivaria.ca/index.php/archivaria/article/view/11758

Fahmy, E., & Bell, K. (2017). Using paradata to evaluate survey quality: Behaviour coding the 2012 PSE-UK survey. In R. Edwards, J. Goodwin, H. O'Connor, & A. Phoenix (Eds.), *Working with paradata, marginalia and fieldnotes, edward elgar publishing* (Chap. 3, pp. 40–60). https://doi.org/10.4337/9781784715250.00009

Fernández-Fontelo, A., Kieslich, P. J., Henninger, F., Kreuter, F., & Greven, S. (2020). *Predicting respondent difficulty in web surveys: A machine-learning approach based on mouse movement features*. https://doi.org/10.48550/arXiv.2011.06916. http://arxiv.org/abs/2011.06916

Holland, J. (2000). *Emergence: From Chaos to Order*. Popular Science/Oxford University Press. https://books.google.com/books?id=VjKtpujRGuAC

Kreuter, F. (Ed.) (2013). *Improving surveys with paradata: Analytic use of process information.* Wiley Series in Survey Methodology. Wiley.

O'Connor, H., & Goodwin, J. (2017). The secondary analysis of fieldnotes, marginalia and paradata from past studies of young people. In R. Edwards, J. Goodwin, H. O'Connor, & A. Phoenix (Eds.), *Working with Paradata, Marginalia and Fieldnotes, Edward Elgar Publishing* (Chap. 6, pp. 94–114) https://doi.org/10.4337/9781784715250.00012

O'Keefe, C., Brown, J., Gandhi, V., Stiber, M., Sorvik, M., Salvatore, T., Dukart, K., Sarcevic, R., Kim, J., Martinez, J., McIntosh, C., Kate, L., Pal, P., & Rudrawar, A. (2022). *UWB-Biocomputing/Graphitti: Just Before We Break It Again*. https://doi.org/10.5281/zenodo.6525597

Pidd, M. (1999) Just modeling through: A rough guide to modeling. *Interfaces, 29*, 118–132.

Pidd, M. (2007). Making sure you tackle the right problem: Linking hard and soft methods in simulation practice. In *2007 Winter Simulation Conference* (pp. 195–204). https://doi.org/10.1109/WSC.2007.4419601

Pomerantz, J. (2015). Use Metadata. In *Metadata* (pp. 117–132). The MIT Press. https://www.jstor.org/stable/j.ctt1pv8904.9

Schenk, P., & Reuß, S. (2023). Paradata in surveys. In I. Huvila, O. Sköld, & L. Börjesson (Eds.), *Perspectives to paradata — Research and practices of documenting data processes*. Springer.

Simpson, R. (2020). Calling the police: Dispatchers as important interpreters and manufacturers of calls for service data. *Policing: A Journal of Policy and Practice, 15*(2), 1537–1545. https://doi.org/10.1093/police/paaa040

Stiber, M., Kawasaki, F., Davis, D., Asuncion, H., Lee, J., & Boyer, D. (2017). Brain-Grid+Workbench: High-performance/high-quality neural simulation. In *Proceedings of the International Joint Conference on Neural Networks, Anchorage, Alaska*.

Trace, C. B., & Hodges, J. A. (2023). Algorithmic futures: The intersection of algorithms and evidentiary work. *Information, Communication, and Society*. https://doi.org/10.1080/1369118X.2023.2255656

Megan Cohen earned her Masters of Archival Studies from the University of British Columbia's iSchool. She has worked with the InterPARES AI project and the University of Washington at Bothell to better understand the intersections between archival theory, computer science, and emergency communications management. She lives in the US Pacific Northwest.

Jardi Martinez Jordan After receiving a BS in Agricultural Science from EARTH University in Costa Rica, Jardi gained experience in agriculture, operations management, and medical interpreting before transitioning to Software Engineering. He earned a Graduate Certificate in Software Design & Development from the University of Washington, Bothell, and is currently pursuing a Master's in Computer Science & Software Engineering. During his studies, Jardi completed two internships at NASA's Kennedy Space Center, where he developed a framework for functional testing of the Launch Control System's event bus. His Master's thesis focuses on developing mathematical and algorithmic models to simulate Emergency Services Communication Systems (ESCS).

M. Scott Sotebeer is the founder and CEO of Sotebeer Management Ventures, LLC/USA Strategics. He has a Ph.D. in Applied Management and Decision Science and an MBA. He is an advisor in Next Generation 911 for the US Department of Homeland Security's Cybersecurity and Infrastructure Security Agency, the University of Washington NG911 AI research lab, and the NSA's Pacific Northwest Smartgrid Cybersecurity project. He is a member of InterPARES AI, providing first responder and emergency communications expertise. He has over 25 years in first responder management, government affairs, and emergency communications.

Michael Stiber is a Professor of Computing & Software Systems at the University of Washington, Bothell. Prior to that, he was an Assistant Professor at the Hong Kong University of Science & Technology and a Research Assistant Professor at the University of California, Berkeley. He has also been a Visiting Associate Professor at the University of Florida and a Fulbright Scholar in the Institute of Physiology of the Czech Academy of Sciences. He earned his Ph.D. in Computer Science from the University of California, Los Angeles. His research lies at the intersection of computer science and complex systems.

The Role of Paradata in Algorithmic Accountability

Ciaran B. Trace and James A. Hodges

Abstract

This chapter examines how the doings of the algorithm (instantiated through its operations, actions, and steps) and its accompanying algorithmic system are revealed and explored through an engagement with the paradata created as a part of this data-making effort. In doing so, the chapter explores how the concept of paradata helps us understand how information professionals and domain stakeholders conceptualize accountable algorithmic entities and how this influences how they emerge as documented and describable entities. Two complementary frameworks for capturing and preserving paradata for accountability purposes are examined in the process. The first is associated with diplomatic theory and archival notions of context and focuses on the role of paradata for algorithmic transparency. The second is related to knowledge management and to efforts in the AI community to use paradata to create unified reporting models that enhance the explainability of algorithms and algorithmic systems. The chapter concludes by demarcating examples and different use cases for paradata for accountability purposes and the mechanisms by which these agents of transparency and explainability can connect with interested and vested audiences.

Author "Ciaran B. Trace" has died before the publication of this book.

J. A. Hodges (✉)
School of Information, San Jose State University, San Jose, CA, USA
e-mail: james.hodges@sjsu.edu

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_11

1 Introduction

With a surge of new technologies leading to a growth in volume, type, and complexity of data-making efforts and their associated data outputs, steps are underway to discern the concurrent processes of data genesis, maintenance, and use. This chapter examines the algorithm and the system surrounding it as a form of information object in a way that adds to the critical inquiry into the data-making efforts inherent in their construction and use. Drawing on scholarship from information science and the broader research community adopting a critical stance on the study of algorithms and algorithmic culture, the chapter examines how the doings of the algorithm (instantiated through its operations, actions, and steps) and its accompanying algorithmic system are revealed and explored through an engagement with the auxiliary data—or paradata—created as a part of this data-making effort.

Before examining paradata in this context, the central part of this introduction establishes a broader narrative on the meaning of algorithms and the reasons for their critical study as part of today's complex data-making landscape. Indeed, this complexity is corroborated by the fact that data-making in governmental, organizational, community, and personal contexts now includes data traces generated through mobile apps, wearables, the Internet of Things (IoT), social media platforms, robots, and the blockchain (Wolf & Blomberg, 2020; Trace & Zhang, 2021; Desjardins & Biggs, 2021; Nasir et al., 2019; Mohammad et al., 2022). As a complex digital object comprising more than just content, software also exists as a ubiquitous form and output of data-making, of which algorithms (the artifact of interest in this chapter) are among the most studied kinds. As objects, methods, and tools for getting things done, algorithms exist in written and performed forms, defining the steps that computer programs follow to solve matters in the process manipulating and organizing data in some manner. In algorithmic decision-making, data is used to model aspects of the world. Using datasets as inputs to a model involves processing source data into select features or variables that allow for and are relevant to the predictions to be made. As a dominant type of AI technology, machine learning algorithms "operate over data inputs and learn from them in that they refine and develop their representations of the world (their models) in such a way that they can predict outputs based on new inputs, classify inputs, and infer hidden variables" (Mooradian, 2019).

As agents in the world, algorithms are obviously consequential, existing to extract, search, filter, classify, recommend, prioritize, and make predictions about people's identities, preferences, and behaviors, feeding into decisions about hiring, criminal sentencing, credit scoring, financial lending, and the like. Given the import and impact of their reach, efforts are underway to demarcate requirements for creating and consuming AI systems to reduce any likely individual or collective risk posed by these technologies (Castelluccia & Le Métayer, 2019; Piorkowski et al., 2020). The accompanying drive to understand AI systems is such that increased importance and visibility are given to the various types and forms of information (including paradata) that document this data-making activity across the management, design, development, testing, implementation, and deployment phases. These documentation efforts aim at audiences whose decision-making is supported or affected by AI systems as well as system auditors and reviewers (Information Commissioner's Office & Alan Turing Institute, 2020).

Of relevance to this chapter is the fact that the reach, power, and complexity of AI models have solidified calls for organizational accountability around the data-making activity that forms the algorithm, accountability being "an overarching principle characterised by the obligation to justify one's actions and the risk of sanctions if justifications are inadequate" (Castelluccia & Le Métayer, 2019, p. III). More precisely, accountability overlays several essential requirements for AI models and systems.[1] Understandability is acknowledged as a critical extrinsic requirement (property and method) for accountability, highlighting the need for comprehensible information to be provided to interested parties regarding the link between the inputs and outputs of AI systems (Castelluccia & Le Métayer, 2019). In turn, the foundational components for understandability exist in the form of transparency and explainability (Castelluccia & Le Métayer, 2019).

The call for transparent AI is bound to the need for organizations to comply with legal and regulatory frameworks for AI systems, and, in this instance, it is policy documents along with operational records such as code, design documentation, model parameters, and learning datasets that need to be available for scrutiny for transparency to be present. As a "ways of working" approach to documenting AI systems, this method provides oversight and insight into internal operations, "demonstrating that you have followed good governance processes and best practices" throughout your design and use of algorithms (Information Commissioner's Office and Alan Turing Institute, 2020). Explanations break into operational, logical, or causal types, generally created and applied to the algorithm or some local and specific result (Castelluccia & Le Métayer, 2019). The call for explainable AI is tied to the benefits that accrue to organizations in cementing external trust with stakeholders, communities, and individuals by way of increased knowledge and awareness as subjects and consumers of AI systems and services (Information Commissioner's Office and Alan Turing Institute, 2020). In this instance, ex-ante and in medias res analyses and post-hoc reflections need to be available for explainability to be present. Such an outcome-based approach to documenting AI systems involves, for example, "clarifying the results of a specific decision"—that is, "explaining the reasoning behind a particular algorithmically-generated outcome in plain, easily understandable, and everyday language" (Information Commissioner's Office and Alan Turing Institute, 2020, p. 22).

In the remainder of this chapter, we explore how paradata, as a type of information object, helps give further substance to the notion of algorithmic accountability and its associated concepts of understandability, transparency, and explainability.

[1] Fairness is demarcated as a key intrinsic requirement, bringing with it the notion that the training data of an algorithmic decision system should be free from bias (Castelluccia & Le Métayer, 2019).

In particular, through a review of the extant literature, the chapter examines how information professionals and domain stakeholders conceptualize accountable algorithmic entities and how this influences how they emerge as documented and describable entities. Two complementary frameworks for capturing and preserving paradata for accountability purposes are examined in the process. The first approach is related to diplomatic theory, which is an investigative tool used to understand the universal characteristics of archival documents. It focuses on the role of paradata for algorithmic transparency and incorporates archival notions of context. The second is related to knowledge management and to efforts in the AI community to use paradata to create unified reporting models that enhance the explainability of algorithms and algorithmic systems. The chapter concludes by demarcating examples and different use cases for paradata for accountability purposes and the mechanisms by which these agents of transparency and explainability can connect with interested and vested audiences.

2 Professional Considerations and the Concept of Paradata

Paradata is a core construct in information studies research that seeks to capture (literally and figuratively) the means and the mechanisms by which a body of information comes to be. Huvila et al. (2021) clarify that what paradata documents and describes are practices and processes. In a work context, practices encompass resources that manifest as part of pursuing an ongoing and overarching goal or interest. At the same time, processes are put in place to get things done. Processes consist of circumscribed activities or steps carried out using sequential or chained actions (both physical and mental) coupled with appropriate methods, technologies, etc. The result of practices and processes is a defined outcome and an accompanying documentary trail, with paradata functioning as a mechanism to ascertain, model, investigate, contextualize, and reconstruct these past occurrences.

In studying the concept of paradata, the question arises as to why it is necessary to document the practice and process by which information is created and used. In one scenario, paradata arguably serves as a tool for organizations to optimize the business practices and processes from which paradata emerges, providing the knowledge necessary to serve as a feedback loop to create improvements, such as system efficiency and effectiveness. This scenario aligns with the understanding and goals of the knowledge management profession.

As a discipline, knowledge management (KM) focuses on the value that information, whether held personally or in formats that allow it to be shared and exchanged, provides in organizational settings (Williams, 2006). Work by Williams (2006) draws from applied linguistics and semiotics to help delineate the practical and theoretical understandings that frame KM, including how the nature of information is understood within organizational contexts. In this articulation, ante-formal information is “flexible, dynamic, and variable,” while formal information is understood as created and explicitly constructed for means of exchange. As Williams notes, formal information is “the outcome of the strategic choice to forgo some of the play and slippage of everyday language, in order to transcribe and transform particular aspects of everyday conversation into formal information" (Williams, 2006, p. 96, 85). The formalization of information in the context of doing business comes about as part of the ways of "doing things" and "making things" (resulting in information about practices and processes), and of describing the type of context in which they may be used (Williams, 2006, p. 85). As Williams articulates, "these artifacts are, at the most obvious level, physical artifacts, but they can range from simple physical artifacts through the range of natural language, right up to complex computer programs for running, supporting and managing all sorts of processes—both physical and social" (Williams, 2006, p. 85).

The role of the knowledge manager is to help organizations adopt an integrated approach to acquiring, communicating, and utilizing information so that it can be put to optimal use for dynamic learning, situational awareness, problem-solving, decision-making, strategic planning, cost savings, and the like. The KM emphasis is thus on helping people comprehend and gain valuable insights from what is considered an essential resource and asset, no matter its level of formalization or stasis. From the perspective of studying algorithmic systems, the role of paradata within a KM lens is scoped such that tangible and intangible work products (the results of work that include design documentation such as flowcharts, training and learning datasets, internal technical code documentation, etc.) can be utilized by the creator and the user of the algorithm as part of improving and optimizing the data curation and computational processes that feed algorithmic systems. Given the push for algorithmic explainability through "post-hoc interpretability" (Castelluccia & Le Métayer, 2019), paradata (such as documentation about the features/variables and other assumptions used in the design of the algorithm) could also be used in a KM framework to illuminate and impart information about the robustness and logic of the algorithmic process, including helping to explain its associated inputs and outputs (results). In this manner, paradata helps ensure that deployed systems are comprehensible to businesses and other users.

This job of facilitating the subsequent scrutiny of and judgment about practices and processes, including their associated inputs and outputs, is an equally important role for paradata. In this scenario, the problem that paradata solves is tied to the need for AI system accountability via transparency. The information professions of records and information management and archival science are best positioned to support paradata's role in this scenario. Like KM, these professions are attuned to notions of value, but information's nature and significance are understood differently. These recordkeeping professions work from the assumption that what we, as a society, do now and have done in the past can be recorded in a manner that can serve as ongoing evidence of, and thus render an account of, what has happened and why. As Hurley notes, records are "especially relevant in documenting the event that triggers the accountability process, and the action or situation under review" (2005, p. 228). The focus here is squarely on information (paradata) that has been formalized, with a record defined as "a document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference."[2]

Yeo describes records as persistent or enduring representations of occurrents, occurrents being "phenomena that have, or are perceived to have, an ending in time" (Yeo, 2018, p. 130). The entities that records represent include events, activities, transactions, and "states of affairs," defined as how things existed at specific points in time (Yeo, 2018). Marrying diplomatic theory (which we will get to in a moment) with Searle's theory of speech acts, Yeo notes that records also represent "assertive, directive, commissive or declarative acts, which are performed by virtue of a record at the moment of its issuance" (2018, p. 152). This frame provides a view of records in which they are understood as stating propositions and how things are in the world, making inquiries or creating future obligations, undertaking to do or carry out something, and bringing about change by declaring it to be so (Yeo, 2018). Data that is contextualized and that is configured to provide appropriate levels of persistence are also considered records. In this instance, contextualized data is a form of what Yeo calls "assertive records": "representations of statements or assertions that have been made about people, organizations, places, events, the results of investigations or the state of the world" (Yeo, 2018, p.145). In this telling, records denote and attest to personal, organizational, and governmental action and are thus evidence of what people and systems engage in as part of the ongoing conduct of work. From the perspective of the study of algorithms, recorded information enables, instantiates, documents, describes, and serves as evidence of the practices and processes that come into play as part of the decision to deploy advanced computers and applications to specific problems. Paradata, thus, is married to the notion that packaged data in the form of descriptions and documentation are contextualized understandings of work practices and processes.

Algorithmic paradata and the broader world in which this data-making effort takes place are more fully conceptualized by applying insights drawn from diplomatic and archival theory. Information doings have long been pertinent elements of concern for archival science. In particular, monitoring and capturing conceptualizations of practices and processes find a home in archival notions of context. Context comes into play as archivists assume the role of information broker. As an infrastructure, the archive serves as a conduit between creators and subsequent users of historical records, allowing these information objects to settle permanently in place with a guarantee of continued authenticity and usability (Trace, 2022a, b). As part of the work of transporting information across time and place, archivists seek to transcribe the context of its original production and use. In doing so, archivists document the "biography of the records, their creator and creation, the serial processes and activities that brought them into being, and the acts of sedimentation that settle them in systems, all the while seeking to reconstruct this life history within an archival fonds" (Trace, 2020, p. 92). To unpack the constituent parts that contextualize organizational records, archivists also rely on diplomatic theory to

[2] InterPARES 2 Project Glossary, http://www.interpares.org/ip2/ip2_term_pdf.cfm?pdf=glossary

better understand the phenomena at play. Diplomatics offers theories to understand and critique the record and its associated practices and processes.

If in Library and Information Science (LIS), descriptive bibliography entails the close examination and cataloging of a text as a physical object, diplomatics emerged as an analytical technique dating from the seventeenth century to study the authenticity and provenance of recorded information (Duranti, 1998). Now updated to study digital records and recordkeeping systems, diplomatic theory reveals how records emerge from administration by unpacking their foundational and necessary elements. In effect, diplomatics allows us to explore paradata retrospectively while pulling us into the circumstances in which it was created in the first instance. To do so, diplomatics instructs us, entails grappling with a broad recordkeeping system composed of a juridical system, an act, a will (to manifest the act), persons, procedures, and a documentary form.

A juridical system is any circumscribed entity, such as an organization or industry, with rules that bind its members' behavior (Iacovino, 2005). Tied to notions of governance and regulation, juridical systems establish the boundaries wherein records have authority and from which legal and moral obligations can be ascertained (Iacovino, 2005). Within a juridical system, an act constitutes the reason records are brought into being, with records associated with the moment of action in which they partake. A will to manifest the action (what is done for a purpose) is effected through a procedure that, according to diplomatics, consists of the body of written or unwritten rules created to carry out an activity. The procedure brings acts or actions out in the world into the record. Processes are the series of motions by which a person prepares to carry out the acts involved in a procedure. Diplomatics tells us that pointers or clues to procedural contexts may be evident in the substance of the document's text or may leave a documentary residue in the form of annotations (or additions to the record's content) added as elements of intellectual form during various procedural moments.

Different procedure phases are also associated with different types of records and determine aspects of their documentary residue. As modern diplomatics has established (Duranti & Thibodeau, 2006) and the policies of the US National Archives (2020) attest, the algorithm itself constitutes a record, albeit one with no traditional (paper-based) counterpart. In this instance, an algorithm is considered an enabling record that uses its digital form to guide the execution of processes. As Duranti and Thibodeau note, software can be viewed as a record in contexts in which it is "generated and used as a means for carrying out the specific activity in which it participates and stands as the instrument, byproduct, and residue of that one activity" (2006, p. 60). What also ensures that this documentary form rises to the level of a record is that it is "properly maintained and managed as intellectually interrelated parts of records aggregations" (2006, pp. 60, 67).

Overall, the practices, procedures, and processes from which the record is created are noted as a describable context of genesis that results in values and actions out in the world being brought into the record, whether in sequence or in parallel to an action. In effect, what diplomatic analysis allows us to abstract or make visible are critical aspects of administrative activities and action—highlighting the practices and routines that typically govern records creation and their flow throughout an organization. In addition, diplomatics allows us to demarcate those records, and aspects thereof, that are a direct residue and thus evidence of procedural and processual action. In our example, diplomatics highlights that form follows function, with algorithmic code, for instance, a manifestation of organizational priorities, including that of the creators and the practical activities and professional roles they inhabit. The documentary form of the resultant algorithmic code reflects a truism; different parts of the world end up in distinct parts of the record.

3 Further Unpacking Algorithmic Practices and Processes

In revisiting the notion that paradata (in whatever recorded form) document and facilitates subsequent interpretation of and judgment about algorithmic practices and processes, this section delves more deeply into what paradata is necessary to adequately convey the essential parameters that have gone into the production of algorithmic systems for the purpose of accountability. Here we show that extant evidentiary records of AI systems and processes (see *paradata for transparency* in Table 1) provide the documentary fodder to augment existing explainable AI reporting frameworks (see *paradata for explainability* in Table 1) together forming a viable basis for emerging documentation standards in the accountable AI sphere (see Lena Enqvist, 2023).

Beginning with the recordkeeping realm, the literature allows us to frame what paradata should be captured and preserved if the target is algorithmic transparency. To do so involves grappling with the difficulties in defining an AI record and what it means to permanently capture and preserve it from an evidentiary perspective (Mooradian, 2019). As Andresen adroitly notes, “There is no universal method for referencing algorithms, or for telling exactly where a specific algorithm, that is supposed to be the unit to be explained, starts or ends within a system in operation” (2020, p. 135). As Andresen also explains, “records that are generated from automated or algorithmic processes do not necessarily differ much from manually created, captured and organized records in matters of evidential value and trustworthiness,” yet “explaining the content of such records may be more difficult” (2020, p. 129). In claiming that “sufficient explanation cannot be obtained from studying process flow or computer program code alone,” Andresen draws attention to the fact that complex systems (particularly those that draw from dynamic and often volatile data sets using ML or probabilistic outcomes) often generate records that can be “difficult to explain, trace, or recalculate after the fact” (Andresen, 2020, p. 130).

In extrapolating what might constitute a sufficient AI record, Mooradian covers familiar grounds in defining it in terms of the “actions, transactions, and events that are carried out (fully or in part) by AI algorithms” (Mooradian, 2019). In providing examples, Mooradian (2019) makes a case for both practice and process documentation, noting that such materials will likely include policy documents, technical documentation on the algorithm and data used as inputs, base systems

Table 1 Unified Framework of Paradata for Accountability

Paradata for Accountability	
Paradata for Understandability—Transparency	
Examples	*Primary target audience—internal and external (information governance and accountability)*
Management and Planning: Policy documentation and communications, internal and external review and audit documentation and communications, business case documentation, request for proposals (RFPs), procurement documentation and communications, contracts, and sales and licensing agreements.	Developers, suppliers, implementors, auditors, regulators, researchers
Requirements, Design, Prototyping, Development, and Testing: Project documentation, team communications, training and evaluation datasets and associated records regarding data selection and preparation including the data cleaning process (e.g., requirements specifications, design documentation, testing reports, task instructions for data workers), information stored in feature stores (e.g., list of features and their definitions and metadata), data catalogs and dictionaries, training data input instruments (e.g., questionnaires, worksheets, score sheets), developer documentation (requirements, specifications, technical designs, testing), source code, data science notebooks, information stored in model registries (description of models and model owners, input parameters and model performance metrics, references to code, copies or snapshots of training data sets, serialized model artifacts, etc.), and output data.	
Operations, Deployment, and Maintenance: Documentation of impact assessments performed, operating manuals, training materials, documentation of impact assessments conducted (algorithm impact assessments—AIAs), and logs while the AI system is operating.	

(continued)

Table 1 (continued)

Paradata for Accountability		
Paradata for Understandability—Explainability		
Examples	*Purpose*	*Main Target Audience—internal and external (information explainability and exchange)*
Explainability Fact Sheets	Understand and compare new and extant explainability approaches for predictive systems.	Designers, implementors (software engineers), and users of explainable methods (AI and ML practitioners and researchers); regulators and certification bodies.
FactSheets for AI Services	Understand AI services' intended use, performance, safety, security, and provenance.	Producers (data scientists) and consumers/subscribers (other developers) of the AI service.
Model Cards for Model Reporting	Evaluate performance characteristics, design, adoption, and effects of trained machine learning models via metrics that capture bias, fairness, and inclusion criteria.	AI and ML practitioners, model developers, software developers, knowledge managers, organizational adopters, regulators, policymakers, and impacted individuals.
Datasheets for Datasets	Standardize how to document and reflect on the provenance, creation, and use of machine learning datasets utilized in developing commercial AI services and pre-trained models to avoid discriminatory outcomes.	ML researchers and practitioners, dataset creators (product teams—including designers, engineers, etc.), dataset consumers, and other interested stakeholders, including policymakers, journalists, consumer advocates, and academics/researchers.
Data Statements	Allow linguistic datasets (collections of speech, writing, and annotations) to be characterized in ways that mitigate pre-existing and emergent bias in natural language processing (NLP) technologies.	Natural language processing technologists (researchers and system developers) and tech policymakers.
Dataset Nutrition Label	Improve AI systems' fairness, accuracy, and transparency by allowing training datasets to be interrogated in terms of their quality, viability, fitness for purpose, etc., before and during AI model development.	Data specialists (e.g., dataset creators/owners, data scientists, product managers).

design, and testing records, along with the forms of compliance documentation being called for within the AI policy sphere. Andresen (2020) also makes a case for practice and process documentation, suggesting that explanations that shed light on the algorithm and its output must be drawn from both. Thus, one source is from discreet external business practices and associated policy documents that provide the context necessary to shed further light on procedural matters. The other source is specific internal transactional processes, operations, and activities from which additional records emerge. If adequate control is available on the data input, Andresen (2020) notes that operational and policy records should be able to capture explanations of algorithmic outputs that are certain, while only policy records are likely to be able to render explanations of scenarios in which different algorithmic outcomes or explanations were possible. Moving forward, Andresen tasks his readers with further ferreting out "what kinds of records, from what kinds of processes, explanations and predictions may reside in" (2020, p. 140). Drawing from the literature and examining AI workflows and online tools for writing AI documentation, Table 1 provides the starting point from which such work can build.

In writing a scoping document on algorithmic decision-making for the European Parliament, Castelluccia and Le Métayer explain that requirements for AI systems can be either "established a priori" (by design) or "checked a posteriori" (using verification) (2019, p. 25). Considering insights from KM and domain experts adopting a critical stance on algorithms, the second framework unpacked here looks at what paradata should be captured and preserved if algorithmic explainability is the target to be deduced. Paradata, in this context, consists of ex-ante and post-hoc documentation scoped for exchange purposes, illuminating and imparting information about the robustness and logic of the algorithmic process, including helping to explain its associated inputs and outputs. Six of the most prominent efforts to generate information about algorithms and algorithmic systems for explainability are considered here, details of which are incorporated into a unified framework below (see Table 1).

Some explainable documentary frames speak to the algorithmic system more broadly. In contrast, others relate to components, including datasets used to train, build, and evaluate models for AI systems and others (artifacts already standard in some areas of the computer industry). In a nod to the former and drawing from the literature on explainable Artificial Intelligence, Sokol and Flach (2020) delineate the parameters of *explainability fact sheets*, a self-reported list of requirements that offers information to parties (including developers) interested in understanding and comparing new and extant explainability approaches (software tools and techniques) for predictive systems, alongside the method itself. Dimensions, reflecting desired properties of explainable approaches, are operationalized as information "desiderata." Desiderata encompass information on functional requirements including the learning task and problem type to which the explanation is tailored, the component (data, models, predications) targeted by the explanation, applicable feature types and classes of models for the explanation, and its relation to the predictive system (ante and post-hoc); information on operational requirements that characterize how users interact with explainable approaches and under what conditions (e.g., provenance, type, and delivery mechanism of the explanation; how the system and explanation interact; intended function, application and audience for the explanation, etc.); information on the properties (usability criteria) of explanations that makes comprehension possible (including soundness, completeness, contextfullness, complexity, parsimony interactiveness, personalization, novelty, actionability); the effect of explainability on the robustness, security, and privacy of the system; and information on any validation measures (user studies or synthetic experiments in settings comparable to deployment scenarios) taken on explainability approaches.

Also proposed in terms of looking at these issues from a functional level are documentation efforts in the form of service-level declarations that emanate from suppliers of AI services to increase confidence in their finished products (in contrast to other efforts described below that focus on datasets or machine learning models). In this scenario, the producers are understood as data scientists, while the consumers of the AI service are pegged as other developers (Arnold et al., 2019). Modeled on industry documents called supplier's declarations of conformity (SDoCs), *FactSheets* are proposed as self-reported information about the supplier, their services, and the characteristics of the development team; the intended domains, purpose, usage, procedures, implemented algorithms, and outputs of the AI service; the methodology and results of associated supplier and third-party safety and performance testing (including which datasets the service was tested on); any potential harms that could result from using the AI service and associated mitigation efforts (including features that relate to fairness, explainability, and accuracy of predictions); security concerns and sensitive use cases; and maintenance of the lineage of the AI service (which speaks to issues surrounding the auditability of data sets and trained models).

With an intended audience of AI and ML practitioners, developers, adopters, regulators, policymakers, and impacted individuals, a frame dubbed "*model cards*" extends the notion of what it means to evaluate how well human-centric AI and ML-trained models perform through the inclusion of metrics that "capture bias, fairness and inclusion criteria" (Mitchell et al., 2019, p. 220). "Model cards" provide a means of disclosing the nature of the model (its who, what, when, and how) and the contexts and domains of use to which it is suited or not suited; model performance across relevant factors, including population groups (cultural, demographic, phenotypic, and intersectional), input instrumentation, and deployment environment; model performance metrics; the datasets used to train and evaluate the model (including how they were chosen and any pre-processing activities carried out on the data); results of the model performance (qualitative analysis) disaggregated by selected factors; and any ethical considerations, challenges, and recommendations noted as part of model development. Providing insights useful for interrogating models' performance, design, adoption, and effects, this documentation is considered a form of ethical reporting intended to be used alongside reporting methods for datasets (Datasheets, Nutrition labels, Data Statements, Factsheets, etc.).

The *datasheets for datasets* frame is built to address a need for a standardized way to document ML datasets to "increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks" (Gebru et al., 2021, p. 86). The aim is to convey information from creators (e.g., product teams) to consumers and other interested stakeholders, including policymakers and academics (Gebru et al., 2021). The datasheet template promotes ex-ante reflection and post-hoc recording of information attuned to the dataset lifecycle or workflow: motivation, composition, collection process, pre-processing/cleaning/labeling, uses, distribution, and maintenance. Scoped for natural language processing systems, *data statements* are similarly envisioned as a new feature of professional practice, in this case, one that allows for linguistic datasets (collections of speech, writing, and annotations) to be characterized in ways that "provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software" (Bender & Friedman, 2018, p. 587). Viewed as a necessary extension of the NLP field, the goal is to have long, or short-form data statements accompany publications on new datasets and experimental results and be included in NLP system documentation. Ideally created contemporaneously with dataset creation, the information schema for data statements consists of the curation rationale for texts; language variety; speaker, annotator, and curator demographics; speech situations; text characteristics; recording quality; and dataset provenance.

A final exemplar is the *dataset nutrition label,* a prototype diagnostic tool (consisting of a method, an associated documentary process, and an interactive web-based application) aimed at improving the fairness, accuracy, and transparency of AI systems by allowing training datasets, or proxies thereof, to be interrogated in terms of their quality, viability, fitness for purpose, etc. before and during AI model development (Holland et al., 2018; Chmielinski et al., 2022). The documentation that makes up a dataset nutrition label aggregates and distills essential information for use by data specialists to inform conversations about dataset quality, specifically their fitness for statistical use cases. Established in a modular fashion, the various prototypes of the tool have incorporated data that are technical and non-technical in nature, with modules generated manually from as-is information (e.g., meta-information about the dataset, information regarding data provenance, and textual descriptions of variables in the dataset) and, unlike the datasheets example, by automated statistical processes to find patterns, relationships, or anomalies (e.g., information about dataset attributes via summary statistics, visualized pair plots, and heatmaps of ground truth correlations).

4 Discussion and Conclusion

Now that the nature and form of paradata for accountability have been examined in the AI sphere, the question turns to the mechanisms through which these agents of transparency and explainability can connect with interested and vested audiences, experts, or otherwise. One idea floated for AI services is to have suppliers post and distribute explainability documentation like *Factsheets* via the blockchain (Arnold et al., 2019). In the local government sphere, public AI registries are touted as a mechanism for people to have understandable and up-to-date information about AI systems (Ada Lovelace Institute, 2021). These online database registries—adopted by city governments in Canada (Ontario), Finland (Helsinki), and France (Antibes, Lyon, Nantes)—are created to capture and make available information from suppliers on the purpose, responsible parties, datasets, data processing, impacts, oversight, and mitigation measures for individual AI systems (Meeri et al., 2020).

In this and other instances of explainable AI, the capture and subsequent use of paradata are situated within an accountability framework in which the populace (those in civil society) is provided with a contemporary window into how and why AI is being used as part of governing structures and activities, with the ability to understand and thus question its benefits and limitations. In moving forward with such registries, additional work will be needed to determine how to integrate appropriate documentation into organizational AI development practices and processes (including ascertaining the responsible parties for explainable paradata creation), as well as how to provision the paradata in terms of scope and detail such that it is fit for purpose (Meeri et al., 2020). Indeed, the pressure to ensure that the use of AI is human-centered (comprising trusted systems and services) has led to the call for at least one associated field (computer vision) to have "dedicated dataset professionals"; professionals undertaking data curation activities in association with external stakeholders and in a manner that aligns trust with "purposefully constructive reporting" (Famularo et al., 2021, p. 2; also see Jo & Gebru, 2020). This chapter contributes to the question of who is qualified to perform this and other human data labor by demarcating how information professionals are already scoped to take on such a role.

As noted in this chapter, the "right to an explanation" approach to accountability is not the only recourse for issues of information asymmetry in the AI environment. As the Ada Lovelace Institute notes, an "under-considered" form of accountability "concerns the preservation and archiving of algorithmic systems for historical research, oversight or audits" (2021, p. 48). As this chapter demonstrates, a complementary form of accountability is possible when records management and archival mandates are put to work to control, manage, and subsequently preserve the paradata necessary to provide transparency about the practices and processes of creators and users of algorithmic systems. Beyond an immediate "need to know" from an internal governance perspective, paradata that can hold people to account provide retrospective and internal transparency (Heald, 2006). As such, it can be utilized by those with a vested interest in auditing and studying the inner workings of the development and impact of AI systems over the long term. Overall, the combination of in situ and post hoc paradata and the requisite skills of information professionals should allow digital registries and archives to function as critical intermediaries between those who create and develop AI systems and those who require or engage in their critical study. In moving forward with AI archives, further work will be needed in the records and information management spheres to review the mandates and regulatory environments surrounding AI practices and procedures and to undertake work process analysis as a prerequisite to developing collection and disposition rules for AI paradata, including identifying what to transfer to an archival repository for long-term preservation. Archivists will also need to supply the curation activities to allow paradata to remain accessible and contextualized, with the specifications for such work currently being investigated in the literature (Van der Knaap, 2020; Hodges & Trace, 2023; Trace & Hodges, 2023).

Acknowledgments This work was supported by Good Systems, a research grand challenge at the University of Texas at Austin.

References

Ada Lovelace Institute, AI Now Institute and Open Government Partnership. (2021). *Algorithmic accountability for the public sector*. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

Andresen, H. (2020). A discussion frame for explaining records that are based on algorithmic output. *Records Management Journal, 30*(2), 129–141. https://doi.org/10.1108/RMJ-04-2019-0019

Arnold, M., Bellamy, R. K. E., Hind, M., et al. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development, 63*(4/5), 1–31. https://doi.org/10.48550/arXiv.1808.07261

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604.

Castelluccia, C., & Le Métayer, D. (2019). European Parliament Scientific Foresight Unit (STOA). *Understanding algorithmic decision-making: Opportunities and challenges* (Report No. PE 624.261). European Parliamentary Research Service.

Chmielinski, K., Newman, S., Taylor, M., Joseph, J., Thomas, K., Yurkofsky, J., & Qiu, C. Y. (2022). The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint, arXiv:2201.03954.*

Desjardins, A., & Biggs, H. R. (2021). Data epics: Embarking on literary journeys of home internet of things data. *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-17. doi:https://doi.org/10.1145/3411764.3445241.

Duranti, L. (1998). *Diplomatics: New uses for an old science*. Scarecrow Press.

Duranti, L., & Thibodeau, K. (2006). The concept of record in interactive, experiential and dynamic environments: The view of InterPARES. *Archival Science, 6*, 13–68. https://doi.org/10.1007/s10502-006-9021-7

Enqvist, L. (2023). Paradata as a tool for legal analysis: Utilizing data on data related processes. In I. Huvila, O. Sköld, & L. Börjesson (Eds.), *Perspectives to paradata - Research and practices of documenting data processes* (pp. xxx–xxx). Springer.

Famularo, J., Hensellek, B., & Walsh, P. (2021). Data stewardship: A letter to computer vision from cultural heritage studies. *Proceedings of the CVPR workshop beyond fairness: Towards a just, equitable, and accountable computer vision*, 25 June 2021.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

Heald, D. (2006). Varieties of transparency. *Proceedings of the British Academy, 135*, 25–43.

Hodges, J. A., & Trace, C. B. (2023). Preserving algorithmic systems: A synthesis of overlapping approaches, materialities and contexts. *Journal of Documentation*. https://doi.org/10.1108/JD-09-2022-0204

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint, arXiv:1805.03677.*

Hurley, C. (2005). Recordkeeping and accountability. In S. McKemmish, M. Piggott, & F. Upward (Eds.), *Archives: Recordkeeping in society* (pp. 223–253). Chandos Publishing. https://doi.org/10.1016/B978-1-876938-84-0.50009-3

Huvila, I., Greenberg, J., Sköld, O., Thomer, A., Trace, C., & Zhao, X. (2021). Documenting information processes and practices: Paradata, provenance metadata, life-cycles and pipelines. *Proceedings of the Association for Information Science and Technology, 58*(1), 604–609. https://doi.org/10.1002/pra2.509

Iacovino, L. (2005). Recordkeeping and juridical governance. In S. McKemmish, M. Piggott, & F. Upward (Eds.), *Archives: Recordkeeping in society* (pp. 255–276). Chandos Publishing. https://doi.org/10.1016/B978-1-876938-84-0.50010-X

Information Commissioner's Office & Alan Turing Institute. (2020). *Explaining decisions made with AI*. https://ico.org.uk/media/about-the-ico/consultations/2616434/explaining-ai-decisions-part-1.pdf.

Jo, E. S, & Gebru, T. (2020, January). Lessons from archives: Strategies for collecting sociocultural data in machine learning. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 306–316.

Meeri, H., van de Fliert, L., & Rautio, P. (2020). *Public AI registers: Realising AI transparency and civic participation in government use of AI*. https://algoritmeregister.amsterdam.nl/wp-content/uploads/White-Paper.pdf

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P. B., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru T. (2019, January). Model cards for model reporting. *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229. doi:https://doi.org/10.1145/3287560.3287596.

Mohammad, A., Vargas, S., & Čermák, P. (2022). Using blockchain for data collection in the automotive industry sector: A literature review. *Journal of Cybersecurity and Privacy, 2*(2), 257–275. https://doi.org/10.3390/jcp2020014

Mooradian, N. (2019). AI, records, and accountability. *ARMA Magazine.*

Nasir, J., Norman, U., Johal, W., Olsen, J. K., Shahmoradi, S., & Dillenbourg, P. (2019, October) Robot analytics: What do human-robot interaction traces tell us about learning? *2019 28th IEEE International conference on robot and human interactive communication (RO-MAN)* (pp. 1–7). IEEE. doi:https://doi.org/10.1109/RO-MAN46459.2019.8956465.

National Archives, Office of the Chief Records Officer. (2020). *Cognitive technologies white paper: Records management implications for internet of things, robotic process automation, machine learning, and artificial intelligence*. https://www.archives.gov/files/records-mgmt/policy/nara-cognitive-technologies-whitepaper.pdf

Piorkowski, D., González, D., Richards, J., & Houde, S. (2020). Towards evaluating and eliciting high-quality documentation for intelligent systems. *arXiv preprint. arXiv:2011.08774.*

Sokol, K., & Flach, P. (2020, January). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Proceedings of the 2020 conference on fairness, accountability, and transparency, USA*. doi:https://doi.org/10.1145/3351095.3372870.

Trace, C. B. (2020). Maintaining records in context: A historical exploration of the theory and practice of archival classification and arrangement. *The American Archivist, 83*(1), 91–127. https://doi.org/10.17723/0360-9081-83.1.91

Trace, C. B. (2022a). Archival infrastructure and the information backlog. *Archival Science, 22*(1), 75–93. https://doi.org/10.1007/s10502-021-09368-x

Trace, C. B. (2022b). Archives, information infrastructure, and maintenance work. *Digital Humanities Quarterly, 16*(1) http://www.digitalhumanities.org/dhq/vol/16/1/000603/000603.html

Trace, C. B., & Hodges, J. A. (2023). Algorithmic futures: The intersection of algorithms and evidentiary work. *Information, Communication, and Society*. https://doi.org/10.1080/1369118X.2023.2255656

Trace, C. B., & Zhang, Y. (2021). Minding the gap: Creating meaning from missing and anomalous data. *Information and Culture, 56*(2), 178–216. https://doi.org/10.7560/IC56204

Van der Knaap, T. (2020). *Honesty through archiving: The contribution of archiving to fair algorithm use by municipal authorities*. [Master's thesis Heritage Studies: Archival and Information Studies (dual), University of Amsterdam].

Williams, R. (2006). Narratives of knowledge and intelligence . . . beyond the tacit and explicit. *Journal of Knowledge Management, 10*(4), 81–99. https://doi.org/10.1108/13673270610679381

Wolf, C. T., & Blomberg, J. L. (2020). Making sense of enterprise apps in everyday work practices. *Computer Supported Cooperative Work, 29*, 1–27. https://doi.org/10.1007/s10606-019-09363-y

Yeo, G. (2018) *Records, information and data: Exploring the role of record keeping in an information culture*. Facet.

Ciaran B. Trace (1971–2024) was a professor at the School of Information at the University of Texas at Austin. She studied the nature, meaning, and function of written records in the lives of those who create and use them. Her research examined the role and meaning of traditional governmental, administrative, and personal records while also examining voluminous data traces generated through algorithmic systems, social media platforms, wearables, and the Internet of Things. Spanning the period from the Progressive Era to modern times, her research illuminated the information worlds of disparate communities, including Southern state archival agencies, 4-H clubs, self-trackers, ovarian cancer patients, law enforcement, digital archivists, and humanities scholars.

James A. Hodges is an Assistant Professor in the School of Information at San José State University. Hodges' research examines the evidentiary value of digital objects, including algorithmic artifacts ranging from conceptual pseudocode to source code, binary files, and elements of sociotechnical context. By closely attending to both the material construction and socially embedded quality of digital artifacts, he brings increased specificity to the analysis of technical objects and creates actionable frameworks for the developers, users, and organizations tasked with creating and managing algorithmic systems. His previously published research applies this approach to the maintenance of biomedical devices, the attribution of uncredited digital labor, and the preservation of historical software.

Adding Paradata About Records Processes via Information Control Plans*

Saara Packalén and Pekka Henttonen

Abstract

To ensure evidence and to control systematically records' life span in digital environment, we need additional data about records' background, history, and actions creating those records. National and international specifications typically set requirements for metadata and functionality that an electronic records management system must have. Creation of metadata is resource consuming. One solution to this problem is to hide and automate records management processes. The chapter examines how this has been done in Finnish public administration. Firstly, the chapter contributes to discussion about description of records management processes and adds understanding of possibilities for adding metadata to records. Secondly, we aim to stir up interest toward the use of a concept paradata in recordkeeping and invite discussion of benefits of understanding some of recordkeeping metadata as paradata. While paradata is not an established term in archives and records management, it is a befitting concept to describe information that is gathered about records during their life span.

*No Creative Commons Licence applies to the respective Third-Party Material.

S. Packalén (✉)
Department of Geographical and Historical Studies, University of Eastern Finland, Joensuu, Finland
e-mail: saara.packalen@uef.fi

P. Henttonen
Faculty of Information Technology and Communication Sciences (ITC), Tampere University, Tampere, Finland
e-mail: pekka.henttonen@tuni.fi

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_12

1 Introduction

A key records management standard ISO 15489-1:2016 defines records management (RM) being "field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposition of records, including processes for capturing and maintaining evidence of and information about business activities and transactions in the form of records" (ISO 15489-1:2016). To ensure evidence and to control systematically records' life span in digital environment, we need additional data about records' background, history, and actions creating those records. In archives and records management transition from paper era to digital environment led to a paradigm shift when archives could no more be managed and considered as physical objects and entities. It was realized that it was not enough to manage existing records only. In digital environment the premise for managing records had to be the process that produced those records. Archival theorist Terry Cook (2001, 4) said that:

> For archivists, the paradigm shift requires moving away from identifying themselves as passive guardians of an inherited legacy to celebrating their role in actively shaping collective (or social) memory. Stated another way, archival theoretical discourse is shifting from product to process, from structure to function, from archives to archiving, from the record to the recording context, from the "natural" residue or passive by-product of administrative activity to the consciously constructed and actively mediated "archivalisation" of social memory.

This shift to processes and functions can be seen in records management metadata. Its roots are in the 1990s. At that time, it had become obvious that recordkeeping professionals (records managers and archivists) must manage also electronic information and abandon their traditional role as custodians of physical documents only (Bearman, 1994; Cook, 1994; Gilliland-Swetland, 2005; Sprehe, 2000). This led to the question of what it means that something is record in an electronic environment and what are requirements for systems that manage electronic records. This was studied in research projects of the University of Pittsburgh and the University of British Columbia (for details, see, e.g., Marsden, 1997). The projects formed the basis for later national and international specifications for electronic records management systems (Gable, 2002; Wilhelm, 2009). The specifications typically set requirements for metadata and functionality that an electronic records management system must have (e.g., the system must capture date and time when information is stored in it and prevent unauthorized destruction and modification of records). Because of the projects it became an axiom in archival science that records systems must link records to business activity/transaction from which they arose (Lappin et al., 2021). This happens by assigning records a place in a functional classification scheme. Today international standard about records management metadata (ISO 23081-1:2017) shows broad consensus about content of metadata.

Creation of metadata is resource consuming. Adding metadata manually is for the person receiving or creating a record often a superfluous step for which there is no motivation, because it usually does not benefit the work task at hand and makes the process slower. One solution to this problem is to hide and automate records management processes. In this chapter, we examine how this has been done in Finnish public administration. Firstly, the chapter contributes to discussion about description of records management processes and adds understanding of possibilities for adding metadata to records. Secondly, we aim to stir up interest toward the use of a concept paradata in recordkeeping and invite discussion of benefits of understanding some of recordkeeping metadata as paradata.

2 Information Control as Part of Information Governance

The aim of controlling information via automated records management processes and metadata serves the goal of information governance (IG) or records/information management which is according to Brooks (2019, 14) "supporting an organization to manage, secure, access and exploit its information in complex digital environments across a myriad of locations."

Today, alongside a relatively narrow concept of RM or records and information management (RIM) more holistic and broader-reaching view of IG has enhanced extensive interest among the recordkeeping professionals. When RM and RIM focus on control of the creation, receipt, maintenance, use, and disposition of records, the concept of IG represents more wide-ranging area of organizations' information needs.

"In short, IG is about information control and compliance" (Smallwood, 2014, 6). Smallwood (2014) sees information governance as a subset of corporate governance. It is about standardizing and systematizing handling of information. It focuses on access, control, management, sharing, storing, preserving, and auditing of information. Organizations' policies, processes, and technologies to manage and control information must be complete, current, and relevant. Further, including "[. . .] *who* is able to access what information, and *when*, to meet external legal and regulatory demands and internal governance policy requirements" (Smallwood, 2014, 6).

Yet, the concept of IG is still vague and there is no one commonly accepted definition of it. High-level nature and breadth of the scope characterizes the various definitions of the concept (Brooks, 2019). A decade ago, Hagmann (2013, 229) stated that "The RIM community tries to capitalize this term [information governance] in order to get a seat at the table of senior executives and to get out of the dusty image of records administration in a paper environment." Lately, other aspects based on the genuine need for broadening the focus in current administrative recordkeeping have also been arisen (Brooks, 2019). The above presented IG definitions of Smallwood pointing to information control fit well in Finnish public sector recordkeeping context in which a proactive recordkeeping strategy that is based on organizational functions has been traditionally dominant.

The Finnish Act on Information Management in Public Administration (906/2019) that was published after and in part as consequence of the GDPR (General Data Protection Regulation) of the European Union has been crucial for the wider understanding of what records management is about. The act made it obvious that in a digital environment one needs a broader approach to information management in which records management is only one part. Although it is possible to manage records without automatic information control, information control allows automatization of records processes and connecting them to information resources of the organization.

In Finland, the National Archives has traditionally played a strong role in guiding public sector organizations' records management. In 2005, National Archives' SÄHKE specification started to stipulate the requirements and features for records' digital archiving in information systems. Finnish SÄHKE2 specification includes a metadata model whose purpose is to ensure evidentiality, integrity, and usability of records (Mäkiranta, 2020). After GDPR and the following regulation, in accordance with the role of National Archives, starting from the beginning of year 2023, SÄHKE2 specification serves only as a recommendation for the agencies.

3 Records Management Metadata and Paradata

Concept of paradata is ambiguous. Current definitions of paradata often include a certain perspective, for example the context of education or research methodology (Pomerantz, 2015), surveys (Kreuter, 2013), or heritage visualization (Baker, 2012). Sköld et al. (2022) discuss paradata in different information domains as well as close connection and overlapping between concepts of paradata, metadata, and provenance data. Only recently the concept of paradata has been introduced to the archival and recordkeeping sphere in studies focusing on paradata in AI-based automation (Davet et al., 2022, 2023). As Davet et al. (2023) state, conceptual overlapping exists between the conceptual development of paradata for AI and those of contextual metadata and explainable AI.

Hence, in archives and records management, the concept of paradata is only emerging to theoretical and practical discussions and thus, it mostly represents an uncharted territory. The concept is barely mentioned in studies in this research area. Studies focusing on metadata or data processes (see, e.g., Bak, 2016; Sundberg, 2013) do not use the term paradata. In similar fashion, Finnish SÄHKE2 calls metadata all data describing the context, content, structure, management, and handling of information (Arkistolaitos, 2008a). Nevertheless, concept of paradata might be applied in this area, too.

In a digital environment, adding metadata is inevitable and an established practice in records' handling and archiving. It is questionable, should we even call it adding, since in digital environment, most of the metadata are automatically added by the recordkeeping system. Some of it are still, though, explicitly added by a human. Metadata is part of the record, part of its content. Meta and data are not to separate any more the way they are/were in the world of paper records (Bak, 2016).

If metadata is defined as information that helps in semantic interpretation of the data, and paradata is all other information about the background, administration, and use of the data, records management metadata belongs almost exclusively to the category of paradata. Although records management metadata may help to interpret the records, it is not generally about the meaning or content of data.

The metadata can be broken into the following components (ISO 23081-1:2017, 16):

1. metadata about the record itself;
2. metadata about the business rules or policies and mandates;
3. metadata about agents;
4. metadata about business activities or processes;
5. metadata about records management processes.

Figure 1 gives an example of the diversity of metadata in archives and records management area. It describes entities in records management metadata. For example, there are metadata about agents, mandates, and business (McKemmish et al., 1999).

In records management, metadata has often a temporal triptych structure: the metadata gives information about current status of records, but also about future and past actions. For instance, metadata may tell that access to records is now restricted, but that the access restrictions will be removed in the future. Once when there are no more access restrictions, the metadata will show what restrictions there have been in the past. This reflects basic conception of records as evidence of past actions.

Metadata accumulates throughout the records' life span. As considered above and shown in Table 1, there are different types of metadata in records management. Metadata is largely about context of records, their background, administration, and use of the records. Some metadata is added at point of capture, that is, when the records are stored in a records management system. After capture metadata is complemented and this continues even when the record has been archived. A study of records in an electronic records management system showed that 65% of metadata was about event history (Kettunen & Henttonen, 2010).

Recordkeeping metadata describes records provenance and relationships that define authenticity, reliability, accountability, and accessibility of digital records throughout the records' life span (Fig. 1 and Table 1). All this data is called metadata. Some of it is various contextual information that is needed to understand the record's provenance and its connections to other records. Much of the data, however, is something else, information about the process and various agents that are related to the record during its life span.

4 Cost of Metadata Creation

Studies have shown that capturing metadata about data context is generally expensive and labor intensive (Faniel et al., 2019). Records management metadata is no exception.

Metadata schemes in records management are broad. For instance, the first version of the Finnish SÄHKE metadata specification includes over 120 elements,

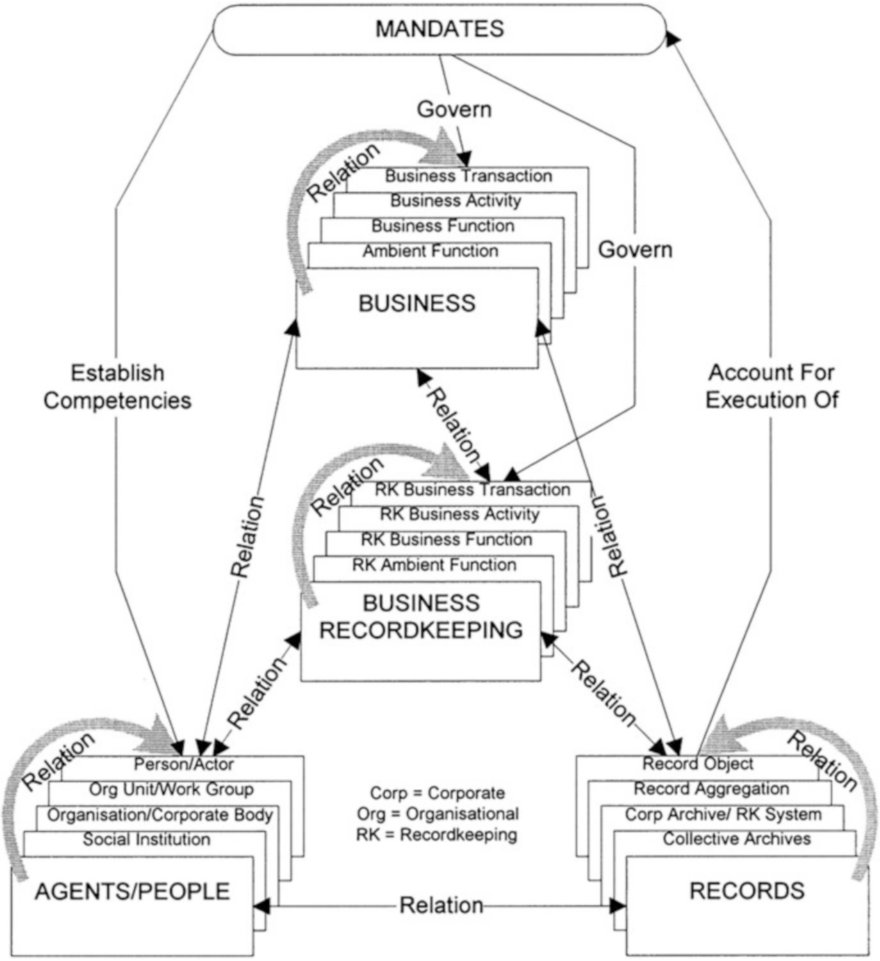


Fig. 1 Coverage of recordkeeping metadata (McKemmish et al., 1999, 15)

many of which can be used at different levels of hierarchy (Records Creator—Collection of records—Record Series—Matters—Transactions—Records). Altogether there are about 280 possible metadata element—entity combinations. A study showed that in one electronic records management system more than half of the metadata elements were unused (Kettunen & Henttonen, 2010). A reason for this may be that while a records management metadata scheme must be prepared for all eventualities when it is concretely applied not all parts of the scheme are necessary. For instance, if the agency does not take part in eGovernment service processes, elements supporting eGovernment services are unnecessary.

The same study showed that optional metadata elements in the scheme were generally ignored, and only mandatory elements had values. Metadata values either

Table 1 Examples about types of metadata in records management. Created from ISO 23081-1:2017

Metadata about	At point of capture	After capture
Records	Date and time when created; record structure; link to business activity or transaction generating the record	Changes in structure or in technical dependencies
Accessibility	Identifiers of records and record aggregations; classification	Terminology changes; changes of personnel; changed locations
Security	Access restrictions	Personnel changes; change of security rules and levels
Business rules, policies, and mandates	Metadata schema; business rules regarding record creation	Metadata showing management of records in compliance with regulatory and other requirements (e.g., access to records)
Agents	Agents involved in record creation	Changes in roles of records
Business process	Information about transactions, but also records management processes (e.g., disposition metadata)	Business processes in which records have been used; copying of records

come from the system, they are default values based on user selection, or free-text values given by the user. A closer inspection of the elements suggested that human intervention was minimal: it seems that users avoided inputting metadata (if they had a choice), and they also preferred to accept default values as such (Kettunen & Henttonen, 2010).

Altogether this—that only mandatory values were given, and that they were generally generated by the system—reveals the high cost of metadata creation. The metadata guarantees authenticity, reliability, and usability of records in long run, but generally its generation makes work processes slower and does not benefit the immediate work task at hand. In addition, users who are not professionals in information management or recordkeeping may find it difficult to assign records a place in the organization's functional classification scheme. A Finnish study showed that even experienced professionals face difficulties in using functional classification schemes (Packalén, 2015). Therefore, one possibility is to hide records management processes from the users and automate them as much as possible. This can take place as part of integration of records and business systems which can take place in several ways (see, e.g., DLM Forum Foundation, 2011, 16–18).

5 Principles of Information Control

The conclusion of the research projects of the University of Pittsburgh and the University of British Columbia in the 1990s was that electronic records management requires (among other things) contextualization of records by preserving information about their functional context and relationships between records.

Finnish recordkeeping had elements supporting contextualization already before digitalization. Like in Nordic countries in general, practice of keeping registries has been common in Finnish administration for centuries. In other words, in- and outgoing letters have been marked in a registry book, card file, or database. Registry entry has joined records together, linked them to a common process, and even to a function (if the registry classification scheme is function based). In short, a registry has given information a context. Another noteworthy characteristic of Finnish recordkeeping is that functional approach was adopted as a starting point for records management in the beginning of the 1980s. Thus, many agencies had a functional classification scheme even before digitalization. This classification scheme formed the core of records management plan that every agency was required to have by law, and which listed record types by function and gave instructions to their retention and management. First national specification for electronic records management systems (known SÄHKE1) in year 2005 required that this plan, now in digital form, was the source of metadata for electronic records (Henttonen, 2023). Besides registry information, functional classification scheme was another source of information about the context of the records.

Next phase took place in year 2008 when the National Archives Service of Finland (current the National Archives) introduced a new approach to improve information management processes. According to SÄHKE1 specification records management plan was to be included in the electronic records management system. The plan contained information about functions, and record types that were generated in them, and gave default metadata values for the retention and management of record types. The next phase brought two changes. Firstly, records management plan was now separated to a system of its own (*Information Control System*) and it was complemented with information about process steps that are taken in the function. *Information control* was defined as management of information management process in an information system (JHS 191, 2015). Secondly, the idea was that the plan—now called *Information Control Plan* (ICP)—would be the source for records management metadata across information systems in an agency: when the process goes forward information systems get metadata values from agency's ICP. This is shown in Fig. 2 that gives an example of a process from recordkeeping perspective. Organizations' ICP gives metadata needed in records handling in an organization. These metadata will then be stored in organization's information system.

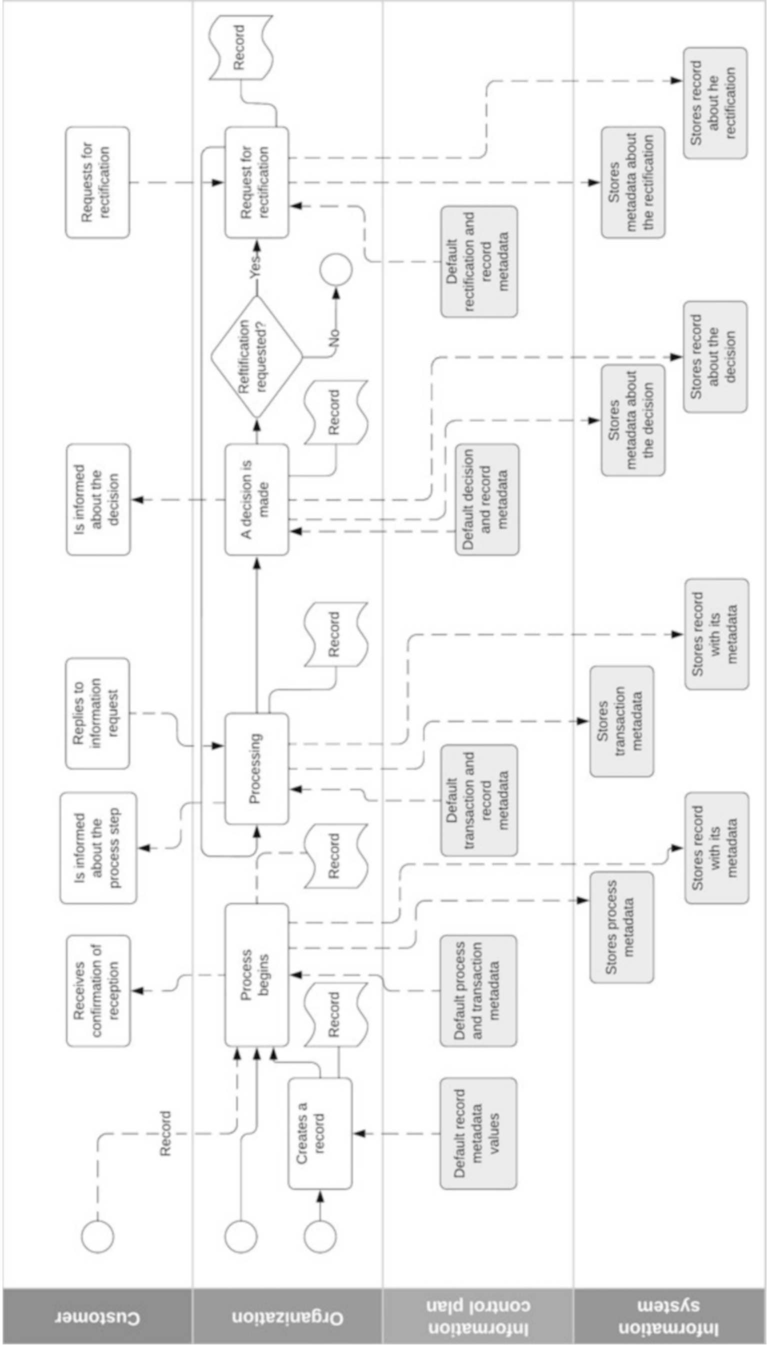

Customer
Organization
Information control plan
Information system
Record
Creates a record
Default record metadata values
Receives confirmation of reception
Process begins
Record
Default process and transaction metadata
Stores process metadata
Stores record with its metadata
Is informed about the process step
Record
Replies to information request
Processing
Default transaction and record metadata
Stores transaction metadata
Record
Stores record with its metadata
Is informed about the decision
A decision is made
Record
Default decision and record metadata
Stores metadata about the decision
Stores record about the decision
Retification requested?
No
Yes
Requests for rectification
Request for rectification
Default rectification and record metadata
Stores metadata about the rectification
Stores record about he rectification
Record

Fig. 2 An example of a recordkeeping process using ICP. Translated from Arkistolaitos (2008b, 3, appendix 1)

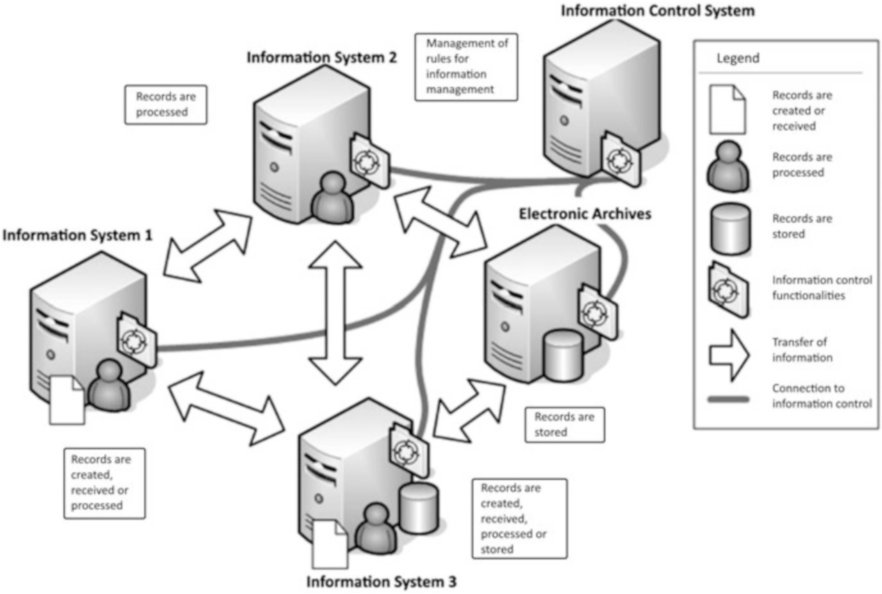

Information Control System
Information System 2
Management of rules for information management
Records are processed
Legend
Records are created or received
Records are processed
Records are stored
Information control functionalities
Transfer of information
Connection to information control
Electronic Archives
Information System 1
Records are created, received or processed
Records are stored
Records are created, received, processed or stored
Information System 3

Fig. 3 Interaction of the Information Control System with other systems. Translated from JHS 176, 2012

Besides SÄHKE specification, legislation about information management, archives, and freedom of information affect information control.[1] There are guidelines and instructions that help to identify and describe business processes in an Information Control Plan. Legislation also defines what information must be included in a registry.[2]

In the core of Information Control System is a plan with an enumerative functional classification scheme which lists all the functions of the agency, and, in addition, process steps and record types that are generated or received in the function. This plan has default metadata values for controlling access to information, managing information security and data privacy, and supporting e-services. Information Control System that contains the plan interacts via Application Programming Interfaces (APIs) with electronic records management systems, electronic archives, and other information systems that process and create records (Kuntasektorin arkkitehtuuriryhmä, 2016). As shown in Fig. 3, Information Control System may control several information systems that, to varying degrees, also interact with each other. Not every information system may store records in an archival system.

[1] Most important laws are Act on Information Management in Public Administration (906/2019), Archives Act (831/1994), and Act on Openness of Government Activities (621/1999).

[2] Act on Information Management in Public Administration (906/2019).

Consequently, in practice, content of an ICP consists of a functional classification describing functions of the organization and descriptions of the organization's operational processes. *Process description* describes an operational process from recordkeeping point of view. It includes administrative (or other) process stages, and transactions that take place in the process phases as well as record types involved. *Process stage* is an entity that includes one or more transactions. In administrative processes there are common administrative process stages (such as initiation, preparation, and decision making) that are similar in every process. *Transaction* is a single task that takes place as a part of a process. JHS 191 recommendation for public agencies (JHS 191, 2015) gives three alternative ways for structuring the content of Information Control Plan. Thus, the plan may state what are

Record types by process stage,

Record types by process stage and supplementary transaction(s) specified by the organization, or

Record types by transactions that are grouped by process stages.

In short, the organization may choose the way it describes their processes in the ICP and how detailed process descriptions it will have. Process descriptions are manually generated data about the processes of the organization. When a process is changed the description must be updated accordingly. Information Control System at hand and agency's other information systems may set limitations to descriptions and dictate what is their appropriate level.

Agencies do not have concrete instructions for creating an ICP besides general level requirements for metadata in legislation and sparse instructions, rules, and regulations given by the National Archives. SÄHKE2 specification used to require (currently it only recommends) describing agency's processes but it does not define sufficient level of detail in the descriptions. This is in the discretion of the agency. It is generally assumed that agencies have already had a functional classification scheme and management data of record types that are generated or received in the functions. This information is in records management plans that precede information control. Thus, if an agency wants to implement information control, it basically only needs to add process descriptions. To create a plan for information control, one needs information from laws and statutes, regulations and standing orders, strategies, quality handbooks, and process descriptions. One needs to consult SÄHKE2 specification and recommendation for the structure of Information Control Plan, JHS 191. One must also discuss with professionals who are responsible for the functions and the processes.

If the agency has analyzed its business processes for other (i.e., business) purposes, those descriptions naturally help in drafting the ICP. However, in an ICP, processes are described from recordkeeping viewpoint. There are no studies about whether and to what extent the result differs from process descriptions that have been created for Business Process Re-engineering, for example. The process in an ICP follows the phases that are carried out in organization's information system,

Table 2 A possible process description of the process of Giving Opinions

Process stage	Transaction	Record type(s)
Commencement	Receipt of the initiating document	Letter/Attachment/Request
Preparation	Processing of the Request for Opinion	Letter/Attachment/Request
	Preparation of the Opinion	Letter/Attachment/Memorandum
Decision making	Drafting/Attaching the Opinion	Opinion/Attachment
	Making a note: No opinion given (note)	Letter
Service notification	Informing officially	Letter

including every possible record type, when handling the matter, e.g., in a recruitment process.

What this means in practice can be seen when we look at administrative procedures which have been the most common targets for information control. For instance, one thing that agencies do is that they give opinions. The ICP has hierarchy of classes in the functional classification scheme:

00 General Administration
00 00 Steering and Development
00 00 02 Opinions
00 00 02 00 Giving Opinions

The lowest level in the scheme is the name of the process. This lowest level groups together process stages, transactions, and related record types. Process stages are common for all administrative processes. Therefore, a process 00 00 02 00 Giving Opinions may be presented in ICP as exemplified in Table 2. Default metadata values governing retention time, access restrictions, etc. of the record types are omitted here. Process stages, transactions, and record types as such are one sort of contextual information (paradata) about records that belong to the process. They help users of the systems to proceed consistently and provide them essential information about the past and forthcoming steps in the process.

Ideally, the agency has SÄHKE2 compatible Information Control System with the appropriate Information Control Plan, the plan is integrated with the electronic records management system and other information systems, and these systems have been adapted to information control. In that case the records process would go as follows:

The user (or personnel in registry office, depending on agency policy) *opens* a new matter in the system and chooses right class for the matter from the functional classification scheme. To support this the Information Control Plan may include additional information (metadata) that helps the user to make the right choice. The user checks the default metadata values and corrects them when necessary. The user may also add supplementary information, like title of the matter, or civil servants handling the matter. Finally, the user *adds* the matter in the registry, and it is assigned a unique registry identification number.

When the Information Control Plan contains process steps, the only path forward allowed is to *follow the process description*. Although the Information Control Plan has default steps, the user may ignore some steps, if necessary, and, e.g., to go from Commencement directly to Decision making, if there is no need for preparatory phases. User cannot add any new steps.

The user then selects the transaction. After selection of the type of transaction the user creates or attaches a *record* to it (when necessary) and selects the appropriate record type from the selection list. The user checks the *default metadata values* and may update them (e.g., define a record that has by default no access restrictions as partly confidential). The user may also add some metadata, like date of receipt, if the system does not supply it automatically. Users' capability to edit metadata is limited by their role. Most users cannot change retention time, for instance. This can be done only by recordkeeping professionals.

When the user creates a new document in the system, the document is first marked as draft, and its visibility is limited. When the document is signed electronically it becomes final and locked to prevent any further changes. The system may include rules that (for instance) mark the matter automatically as *closed* when the process reaches a particular phase.

During the process, two things may happen simultaneously. Firstly, the records gather paradata about their background and the function/process which they are result of. Secondly, this paradata is used as basis for further actions. For instance, if record's retention time is calculated from the completion of the process, i.e., when the matter is closed, date of the process completion is recorded in para-/metadata and used to assign a date when the record is to be removed from the system.

An Information Control Plan can be just a guidebook to agency's functions, processes, and information, but it is stated that full benefit comes only if the plan is used to manage and automate information systems. There are no studies on whether and to what extent this goal of information control has been reached, but SÄHKE2 specification and concept of information control have today established their position in Finnish public sector records management (Mäkiranta, 2020). Other benefits may include improved usability of information systems because of default metadata values and process descriptions (JHS, 2015). Metadata enables tracking of processes and facilitates answering to information requests. Entries in the registry/system document content of a record, processes (when it has arrived, who has created it, etc.) and what transactions (like answering to a request for opinion) have taken place. Information Control Plan not only describes the records accumulating in the course of organizations' business activities but provides a tool for managing processing stages of the information and for information security measures. When fully exploited, several information systems would be controlled by one ICP via APIs. If the systems are used without exceptions, they provide trustworthy evidence of the flow of information in an organization. Persons who access information later may convince themselves about the authenticity and reliability of the information by looking at the para-/metadata about it.

A prerequisite for information control is existence of a functional classification scheme. Functional approach to records management is today widely accepted among recordkeeping professionals in the Finnish public sector (Packalén & Henttonen, 2016a). However, people understand basic concepts, like function differently, and the plans are sometimes difficult to use. Functional approach needs more rigorous theoretical basis (Packalén, 2017). Functional schemes are heterogeneous, and analysis of class names shows ambiguity and varying conceptual structures in the schemes (Packalén & Henttonen, 2016b).

Creating a workable ICP and integration of Information Control System with recordkeeping and other systems requires a collaboration between several stakeholders of the organization: records managers, data protection specialist, lawyers, IT personnel, system suppliers, and specialists on various subject areas like personnel management. In addition, collaboration with National Archives of Finland is necessary to define records with archival value. Once created an Information Control Plan needs constant updating.

Implementation of information control requires financial, technical, and human resources. Even with appropriate resourcing (which is often lacking) the goal is difficult to reach. An unpublished report on electronic archiving in municipalities three years ago revealed that implementation of information control—at least in a rigid form—has not been feasible in most information systems or databases, and that information in the municipalities is generally hybrid and only exceptionally complete digitalization has been achieved (Hänninen, Heli: Sähköisen arkistoinnin tilannekuvan selvitys kunnan toimialoilla 2020). A study in year 2020 found out that only in one state agency out of ten their ICP controlled more than one information system. For some, information control concerned only a part of organizations' functions (Mäkiranta, 2020). No university had entirely digital processes for records with permanent value (Kokkinen, 2020, 10).

6 Discussion and Conclusions

Information control brings together different ideas. Information Control Plans incorporate traditions of Finnish recordkeeping: registry practice, and proactive planning of records' life span. They fulfill internationally recognized requirements for electronic records management and, in addition, serve implementation of Freedom of Information legislation. An Information Control Plan can be accessed by anyone according to Finnish Freedom of Information legislation. Thus, the plan increases transparency by showing what information is gathered and processed in public administration (Lybeck et al., 2006). Information control is a combination of functional approach in records management planning, information governance perspective to information management, local recordkeeping traditions, and need to increase efficiency and automate processes.

SÄHKE2 continues to exist as a recommendation. In the future, The National Archives will primarily focus on records that have value for permanent preservation, archives, and have less authority in records management. For the same reason, in a recent draft for archival legislation there is no requirement for agencies to implement information control (Luonnos hallituksen esitykseksi eduskunnalle arkistolain ja

Kansallisarkistosta annetun lain muuttamisesta, 2022). Thus, formally there will be less constraints and obligations in Finland for public sector records management to fulfil. Nevertheless, proper metadata and goals of information governance are still considered important. While agencies will have more freedom in their records and information management, it is likely that for practical reasons information control and SÄHKE2 specification still form the basis for future development. Although implementations may change, organizations still need to describe their processes and have procedural information about them. Metadata are to ensure later understandability and usability of digital records. Understanding paradata as part of it may bring new, useful insights to the future discussion of process descriptions.

Clearly, information control carried out through an ICP has benefits. Adding metadata (paradata) automatically to records and the processes they originate from accelerates management of information. It shortens the time used in handling a matter in an agency which might lead to increased customer satisfaction. However, there are no studies that would empirically show the benefits of information control. For instance, automation of records processes may save human resources, but there is no research showing how significant these savings are. However, on the other hand, ICP must be constantly kept up to date about information and processes in the organization which is a laborious and resource- craving task. Functional classifications are not without problems. Previous studies have shown several challenges in organizing records by function. Some of them are a result of conceptual confusion and heterogeneous classificatory structures which result from a lack of theoretical background and guidance for creating classifications (Packalén & Henttonen, 2016b). Information Control Plans involve similar challenges.

Information Control Plan as such is a record that is regularly updated. Old and new versions of the ICP are preserved. They provide enormous amount of information about organization's processes and records management. Information Control Plans are not paradata themselves. They are only descriptions of organization's planned functions, processes, and records. When the plans come to flesh in the organization's daily operations, information in the plans becomes paradata about the records. The InterPARES research group defined paradata as "information about the procedure(s) and tools used to create and process information resources, along with information about the persons carrying out those procedures" (Davet et al., 2022). All this information is saved in records management and information control system used.

It is important to understand that when operating in digital environment the premise of the information gathered is the records creating process and not a single document/record. Typically, Finnish recordkeeping practices and procedures are not based on theories but are constructed as resolutions from practical needs (Kilkki, 2004). The same applies to Information Control and its applications. It rests on the rather weak theoretical base of the functional approach to records organization. One should base recordkeeping activities on theoretical and conceptual understanding and underpinnings. Therefore, in archives and records management discipline one needs to examine the potential of the concept paradata from various perspectives. Finding out what it is that paradata has to offer to archives and records management and contributing the recordkeeping viewpoint to paradata discussion is a start. While paradata is not an established term in archives and records management, it is a befitting concept to describe information that is gathered about records during their life span. How we name things forms our understanding. It is about understanding what kinds of data it is that we add to records and the records creating processes, and about understanding the foundation of our actions.

References

Arkistolaitos. (2008a). *SÄHKE2. Sähköisten asiakirjallisten tietojen käsittely, hallinta ja säilyttäminen* [SÄHKE2. Management and preservation of electronic records]. Retrieved April 13, 2023, from https://kansallisarkisto.fi/uploads/normit/valtionhallinto/maarayksetjaohjeet/normiteksti_suomi.pdf

Arkistolaitos. (2008b). *SÄHKE2. Sähköisten asiakirjallisten tietojen käsittely, hallinta ja säilyttäminen. Liite 1*. [SÄHKE2. Management and preservation of electronic records. Appendix 1]. Retrieved April 13, 2023, from https://kansallisarkisto.fi/documents/141232930/154880298/Sahke2-Liite1-Metatietojen_tuottaminen.pdf/3f4f169b-81a2-ed00-2c20-ff0a9feb3731/Sahke2-Liite1-Metatietojen_tuottaminen.pdf?t=1679904728156

Bak, G. (2016). Not meta just data: Redefining content and metadata in archival theory and practice. *Journal of Archival Organization, 13*(1–2), 2–18. https://doi.org/10.1080/15332748.2017.1413974

Baker, D. (2012). Defining paradata in heritage visualization. In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.), *Paradata and transparency in virtual heritage* (pp. 163–175). Taylor & Francis Group.

Bearman, D. (1994). The implications of Armstrong v. Executive Office of the President for the archival management of electronic records. In *Electronic evidence. Strategies for managing records in contemporary organizations* (pp. 118–145). Archives & Museum Informatics. http://www.archimuse.com/publishing/electronic_evidence.html

Brooks, J. (2019). Perspectives on the relationship between records management and information governance. *Records Management Journal, 29*(1/2), 5–17.

Cook, T. (1994). Electronic records, paper minds: the revolution in information management and archives in the post-custodial and post-modernist era. *Archives and Manuscripts, 22*(November), 300–328. http://socialstudies.cartagena.es/images/PDF/no0/cook_electronic.pdf

Cook, T. (2001). Archival science and postmodernism: new formulations for old concepts. *Archival Science, 1*(1), 3–24. https://doi.org/10.1007/BF02435636

Davet, J., Hamidzadeh, B., Franks, P., & Bunn, J. (2022). Tracking the functions of AI as paradata & pursuing archival accountability. In *Archiving 2022: Final Programs and Proceedings, 7-10 June 2022* (pp. 83–88). Society for Imaging Science and Technology.

Davet, J., Hamidzadeh, B., & Franks, P. (2023). Archivist in the machine: Paradata for AI-based automation in the archives. *Archival Science*. Advance online publication. https://doi.org/10.1007/s10502-023-09408-8.

DLM Forum Foundation. (2011). *MoReq2010. Modular requirements for records systems*. Volume 1. Core services & plug-in modules. Version 1.0.

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation, 75*(6), 1274–1297. https://doi.org/10.1108/JD-08-2018-0133

Gable, J. (2002). Everything you wanted to know about DoD 5015.2. *The Information Management Journal, 36*(6), 33–38.

Gilliland-Swetland, A. (2005). Electronic records management. *Annual Review of Information Science and Technology, 39*, 219–253. https://doi.org/10.1002/aris.1440390113

Hagmann, J. (2013). Information governance - beyond the buzz. *Records Management Journal, 23*(3), 228–240.

Henttonen, P. (2023). “One system to rule them all”: The limited success of information control systems in Finland. In B. Greg & M. Rostgaard (Eds.), *The Nordic model of digital archiving* (pp. 115–134). Routledge studies in archives. Routledge.

ISO 15489-1:2016. *Information and documentation - Records management. Part 1: General*. (2016). ISO.

ISO 23081-1:2017 *Information and documentation. Records management processes. Metadata for records. Part 1: Principles*. (2017). ISO.

JHS 176. (2012). *Sähköisten asiakirjallisten tietojen käsittely, hallinta ja säilyttäminen* [Management and preservation of electronic records] Retrieved April 4, 2023, from https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.suomidigi.fi%2Fsites%2Fdefault%2Ffiles%2F2020-06%2FJHS176.doc&wdOrigin=BROWSELINK

JHS 191. (2015). *Tiedonohjaussuunnitelman rakenne* [Structure of Information Control Plan] (pp. 1–17). JUHTA. Retrieved April 4, 2023 from https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.suomidigi.fi%2Fsites%2Fdefault%2Ffiles%2F2020-07%2FJHS191_0.doc&wdOrigin=BROWSELINK

Kettunen, K., & Henttonen, P. (2010). Missing in action? Content of records management metadata in real life. *Library and Information Science Research, 32*(1), 43–52. https://doi.org/10.1016/j.lisr.2009.10.002

Kilkki, J. (2004). Bearmania. Frosting Finnish archival practice with imported archival theory. *Comma, 1*, 43–53. https://doi.org/10.3828/comma.2004.1.7

Kokkinen, S. (2020). *Sähköisen arkistoinnin nykytilan selvitys. Korkeakoulusektori*. [State of electronic archiving. Higher education institutions] Aalto-yliopisto.

Kreuter, F. (2013). *Improving surveys with paradata: analytic uses of process information*. Wiley, Incorporated.

Kuntasektorin arkkitehtuuriryhmä. (2016). *Kuntasektorin asianhallinnan viitearkkitehtuuri* [Reference architecture for records management in municipal sector]. Retrieved February 1, 2023 from https://www.kuntaliitto.fi/sites/default/files/media/file/Kuntasektorin%20asianhallinnan%20viitearkkitehtuuri.pdf

Lappin, J., Jackson, T., Matthews, G., & Ravenwood, C. (2021). Rival records management models in an era of partial automation. *Archival Science, 24*. https://doi.org/10.1007/s10502-020-09354-9

Luonnos hallituksen esitykseksi eduskunnalle arkistolain ja Kansallisarkistosta annetun lain muuttamisesta [Draft for a government proposal about changing the Archives Law and Law about the National Archives]. (2022).

Lybeck, J. et al. (2006). *Arkistot yhteiskunnan toimiva muisti. Asiakirjahallinnon ja arkistotoimen oppikirja* [Archives - the functioning memory of the society. A textbook for records and archives management]. : Arkistolaitos.

Mäkiranta, M.-L. (2020). *Sähköisen arkistoinnin nykytilan selvitys/valtionhallinto* [State of electronic archiving. Government administration]. Retrieved March 28, 2023, from https://okm.fi/documents/1410845/33413091/S%C3%A4hk%C3%B6isen+arkistoinnin+nykytila+raportti+valtionhallinto.pdf/25a91cfa-ed30-639d-99b8-e1fe421a945f?t=1598525687764

Marsden, P. (1997). When is the future? Comparative notes on the electronic record-keeping projects of the University of Pittsburgh and the University of British Columbia. *Archivaria, 43*, 158–173.

McKemmish, S., Acland, G., Ward, N., & Reed, B. (1999). Describing records in context in the continuum: The Australian recordkeeping metadata schema. *Archivaria, 48*, 3–37.

Packalén, S. (2015). Functional classification. Record-keeping professionals’ difficulties and their handling in maintenance and use of FC in Finnish organisations. *Records Management Journal, 25*(2), 166–182.

Packalén, S. (2017). *Functional classification systems in Finnish public-sector organisations*. University of Tampere. http://urn.fi/URN:ISBN:978-952-03-0473-7

Packalén, S., & Henttonen, P. (2016a). Recordkeeping professionals' understanding of and justification for functional classification: Finnish public sector organizational context. *Archival Science, 16*(4), 403–419. https://doi.org/10.1007/s10502-015-9254-4

Packalén, S., & Henttonen, P. (2016b). Ambiguous labels: Facet analysis of class names in Finnish public-sector functional classification systems. *Knowledge Organization: KO, 43*(7), 490–501.

Pomerantz, J. (2015). *Metadata*. MIT Press.

Sköld, O., Börjesson, L., & Huvila, I. (2022). Interrogating Paradata. *Information Research*, 27(Special issue, October 2022). 10.47989/colis2206.

Smallwood, R. F. (2014). *Information governance: Concepts, strategies, and best practices*. Wiley, Incorporated.

Sprehe, J. T. (2000). Integrating records management into information resources management in U.S. Government agencies. *Government Information Quarterly, 17*(1), 13–26.

Sundberg, H. P. (2013). Process based archival descriptions – organizational and process challenges. *Business Process Management Journal, 19*(5). https://doi.org/10.1108/BPMJ-Jan-2012-0002

Wilhelm, P. (2009). An evaluation of MoReq2 in the context of national EDRMS standard developments in the UK and Europe. *Records Management Journal, 19*(2), 117–133. https://doi.org/10.1108/09565690910972075

Saara Packalén PhD, is a University Lecturer at the University of Eastern Finland, in the Department of Geographical and Historical Studies, at the Master's degree programme in Archives and Records Management. After finishing her doctorate from the University of Tampere, Information Studies and Interactive Media, in 2017, she has also worked in Finnish public sector, in project management and recordkeeping duties.

Pekka Henttonen D.Soc.Sc., Adjunct professor, is a University Lecturer in the Faculty for Information Technology and Communication Sciences at the Tampere University. Before academic career he worked in the National Archives of Finland and in the Military Archives of Finland.

Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes

Lena Enqvist

Abstract

This chapter explores aspects of the relationship between technology, transparency, and accountability in public decision-making. It addresses how technological advancements have increased accessibility and automation while complicating decision process reviewability. It explores transparency as a relational concept and focuses on legal obligations on documentation and records-keeping, such as in the EU General Data Protection Regulation and the upcoming EU Artificial Intelligence Act, as a means to bolster transparency and improve reviewability. In particular it also discusses the feasibility of gathering and analysing 'paradata'—data pertaining to data processes—as a means to safeguard legality and transparency in automated decision-making, notably within the public sphere.

1 Introduction

Although various public sector bodies around the globe have been using technologies to assist their tasks and decision-making for decades, such uses have advanced and intensified greatly especially in the last decennary. While these developments hold many promises of highly serviceable and efficient public sectors, concerns have also been raised over the risks of placing too much trust in the technologies' capacities to produce reasoned recommendations or decisions that aligns with the law. Real-life examples, such as the Australian government's use of the so-

L. Enqvist (✉)
Department of Law, Umeå University, Umeå, Sweden
e-mail: lena.enqvist@umu.se

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_13

called RoboDebt system, illustrate how flawed system designs or inappropriate applications can have effects on the legality of the public exercise of power at a large scale. The system was found to have miscalculated hundreds of thousands welfare recipients' incomes and related rights to benefits—and as a consequence automatically issued a vast number of faulty debt collection decisions to citizens often part of socially vulnerable groups (Carney, 2019). Examples like these have contributed to intensified political as well as academic discussions on the effects of technologies on the public exercise of power. At the core of these discussions lies the question of how to ensure transparency and accountability when public power is exercised via technological proxies. That governments are transparent in their exercise of power against the public is, namely, a foundational principle of the 'Rule of law' (Jamar, 2001). Otherwise, the prospect to review whether public powers have been exercised within their limits is hampered. Transparency is, however, not a fixed concept and relates to other similar concepts such as openness, explainability, interpretability, accessibility, visibility, and reason-giving (Felzmann et al., 2019a). There is no obvious or infallible solution to ensuring and safeguarding transparency. This contribution will, however, focus on public authorities' use of technology to assist their decision-making as it has proven to be principally challenging from a public transparency perspective. It will also, in particular, discuss the collection and use of 'paradata' as one possible and advantageous tool and building block of transparency in this context. The formalisation of data on processes that the collection of paradata implies may, namely, prove useful for enabling qualitative reviews of whether automated systems are operating lawfully.

Before continuing to outline this contribution, some introductory definitions are in place. Firstly, 'technologically assisted decision-making' is here used as a broad term including any use of technologies by public authorities to prepare, recommend, or make decisions. This means that fully as well as partially automated decision-making procedures are included. Secondly, 'paradata' is not a legal concept and therefore, naturally, also lacks a legal definition. This contribution will use a wide definition as *any* 'data-on-data related processes and practices' extending beyond its original survey domain (Couper, 2017). Importantly, the definition includes fixed design decisions on system processes, as well as data on how these design decisions have been employed in particular applications. As relevant to the context of public decision-making, 'paradata' here also includes descriptions of the procedural aspects of how a system is designed to run, its authorisation and constraints. Narrow distinctions from neighbouring or partly overlapping concepts such as provenance metadata or contextual metadata, for example, will not be made (Bentkowska-Kafel et al., 2012; Reilly et al., 2021).

The contribution will have the following structure. First, Sect. 2 places the 'law' and legal practice in an information and knowledge management context, and link to how technological developments in legal information and knowledge management have made 'the law' more accessible and automatable at the same time as it has obscured parts of the decision-making procedure and affected its reviewability. Section 3 deepens the analysis by focusing on the general requirements for transparency in public decision-making and combines it with an argument for an increased need to analyse data-on-data related processes, ‘paradata’, as part of ensuring that automated decision-making processes are lawful and transparent. Section 4 then discusses how requirements on documentation and recordkeeping of data-related processes can contribute to increased qualitative transparency in connection with public automated decision-making, focusing on the GDPR and EU’s upcoming Artificial Intelligence Act as examples. In Sect. 5, the merits of legal standards for documentation and recordkeeping on data-related processes are discussed as one important affordance for the utilisation of such data in legal analysis. The chapter is finally concluded through Sect. 6, which discusses the potential benefits and challenges to utilising paradata analysis within the legal domain.

2 Legal Knowledge Management as the Nexus of Legal Practice

Law is a knowledge-based profession and its core, ‘legal practice’, is about providing specialised knowledge, expediated through expert services and the exercise of power (du Plessis & du Toit, 2006). More specifically, legal knowledge concerns the law and its application and is used to produce and manage legal work. Therefore, legal research is and has always been central to any legal practitioner or scholar to find solutions to particular legal questions. This is often a time-consuming task, as legal knowledge is acquired from the internalising of information gathered during legal studies, legal research and legal experience. The primary sources include statutes, preparatory works and case law, etcetera, and the secondary sources include legal reference works, digests, indexes, law reviews, legal periodicals, commentaries, books, and articles from specialised law publications (Roos et al., 1997). Unsurprisingly, functional legal information management is thus imperative to the acquisition and internalising of legal knowledge.

While legal knowledge management traditionally has been intimately tied to human carriers/intermediaries—technologically driven transformations have partially challenged this premise. Early legal information and knowledge management research was primarily concerned with libraries and their roles in aiding legal research through structuring legal information carriers such as statutes, case law or academic writing, etcetera. The introduction of new technologies shifted much of this focus to search engines and the build-up of legal databases to aid the work flows for those performing legal research (Berring, 1994; Foster & Kennedy, 2000). Such systems have functioned as technological drivers for a transformation in the methods that lawyers use to access, retrieve, and process information in order to solve legal problems (du Plessis & du Toit, 2006; Merwe, 1986; Susskind, 2000). The expected promises of the technologically mediated legal information retrieval have been high and span all the way to discussions on whether the new potential efficacy of legal information management might even render lawyers and their tacit knowledge superfluous in some cases (Davis, 2020; Susskind & Susskind, 2016). Because when technology is sufficiently capable of imitating a ‘cognisance’ of the law, there is less need for human intermediaries to translate the law into directives on how to act.

In relation to human decision-makers, automated processes are intended to 'embody the specialised knowledge and experience of a human expert in a chosen domain' and provide a 'mechanism for applying this knowledge to solve problems in that domain' (Kidd, 1985). The aim is, thus, to augment human decision-making through knowledge management. Focusing on automated decision-making or recommender systems, they are more advanced in their operations than performing mere 'legal research'. Their aim is not just to assist in determining what the law is, but also to determine how the law applies in a given situation. They therefore need to combine the (identified as) relevant set of rules with case specific data input, to produce a case specific data output in the form of an individualised decision or recommendation.

A growing credence to automated procedures is signalled through a nearly worldwide general tendency towards national public administrations deploying different types of technology (ranging from simpler pre-programmed 'if–then' statements to more dynamic AI and machine learning applications) to make or support their decision-making. Underlying assumptions are that automation 'done right' will help streamlining decision-making procedures, reducing the need for tacit knowledge, risks of skill-inadequacies and human bias. The expectations also include that automation will be able to provide a more accurate as well as speedy justice for those subject to the public exercise of power, as well as a more cost-efficient administration. Automated decision or recommendation systems are, thus, seen as media for knowledge management to scale up the capacity and effectiveness of knowledge distribution. On the other hand, there are also associated risks. The assuring that public automated processes work satisfactory and lawfully will require that human intermediaries exercise oversight and control of their functioning. Knowledge of how the system operates, and why, is crucial. What information governance regimes are in place thus plays an important role in elevating how the delivery of knowledge occurs to those tasked with overseeing the proper functioning of automated processes. The next section will, therefore, discuss requirements of transparency in relation to automated decision-making, and the possible utilisation of 'paradata' in this context.

3 Transparency in Public Decision-Making and the Growing Need for Analysis of Data-Related Processes

As put by Oswald, algorithmic decision-making may come with the risk of creating substantial or genuine doubt as to why decisions were made and what conclusions were reached (Oswald, 2018). To mitigate these risks, the assisting technologies must serve several and sometimes counterbalancing objectives at the same time. They must not only aid the more obvious aims such as correct and more efficient decision-making procedures. The technologies must also serve several other legal (and 'rule of law') values such as the supremacy of law, equality before the law, accountability to the law, fairness, and legal certainty (Zalnieriute et al., 2019). And, as already introduced, they must also secure sufficient transparency to enable scrutiny of the public exercise of power.

The basic idea is that transparency increases the chances to detect wrongdoings, uncover abuses of power, and scrutinise public activities (Matheus et al., 2021). In this respect, transparency is foremost a supporting value to the realisation of other pertinent values—and essential to establishing trust in the public administration. As put by Jamar, transparency refers to a cluster of related ideas, including governmental action in the open, the availability of information (particularly relating to the law), as well as accuracy and clarity of information (Jamar, 2001). There is no common or comprehensive definition of transparency in the legal sense. Focusing on the aim of transparency, Mock's definition is, however, a useful starting point:

> Transparency is a measure of the degree to which the existence, content, or meaning of a law, regulation, action, process, or condition is ascertainable or understandable by a party with reason to be interested in that law, regulation, action, process, or condition. (Mock, 1999, p. 1082)

Notably, this definition expands from a purely 'informational' perspective on transparency (where open access to data would equal transparency)—into a 'relational' one, which takes the recipient's end into consideration (Felzmann et al., 2020). Transparency is thus not only understood as a quality of being open and overt, but also as a quality of being identifiable and understandable. As the latter aspects indeed depend on the recipient's knowledge base and need for awareness, a key question is for whom the automated decision-making processes are supposed to be transparent (Larsson & Heintz, 2020). This topic is discussed and debated as a matter of achieving 'meaningfully' transparent decision-making (Edwards & Veale, 2017; Felzmann et al., 2019b), where the discussions hook into relating concepts such as 'explainability' (Deeks, 2019) or 'reason-giving'(Ng et al., 2020). Creating and maintaining the transparency of technologically assisted decision-making is therefore, indeed, a process in itself that requires the repeated consideration of the recipients' end of process or data-oriented information.

Complicating the matter is also the fact that not all manifestations of 'transparency' are helpful for the cause of fair and lawful public decision-making. Full transparency into the processes of an algorithmic system may disservice its reviewability by overloading the receiver with information that at least partially requires special expert competence to decode and interpret. The recipient, or overseer, needs to be able to quickly translate the data into knowledge that is useful for identifying whether the system is somehow flawed, and whether a decision or a recommendation is lawful or not. Because even if there were full openness regarding the input data as well as regarding the algorithmic method used, it is primarily the interplay between the two that yields the complexity—and thus opacity (Burrell, 2016).

Transparency can also have negative effects on other legitimate objectives in public decision-making. The limited human control of automated systems makes them susceptible to risks if they are 'too' understandable, as this might open the door to misuse by stakeholders trying to 'game' the system. Particular output objectives of transparency, such as the possibility to identify and correct biases embedded in, or reproduced by, an automated system might contrast with privacy protections (Independent High-Level Expert Group on Artificial Intelligence, 2018; Larsson & Heintz, 2020). Moreover, there are of course also limits to the technical or economic feasibility of providing extensive transparency, as well as limits posed by obligations or wishes to protect intellectual property, trade secrets, national security and defence, as well as public security.

Transparency in relation to algorithmic decision-making is thus complicated. Even so, the obscuring effect that automated processes have in relation to public decision-making makes it clear that the mere disclosure of a system's in- and output is not enough. The fact that the legal rules themselves are public and published is not sufficient to ensure a transparent handling of the input data, as the automated processes will not necessarily interpret or utilise this data in a way that replicates legal reasoning. A focus on the strictly informational aspect of transparency is therefore not enough to ensure efficient scrutiny of public automated or automatically supported decision-making.

Attending also to the relational aspect of transparency requires that the knowledge representation and problem-solving processes employed by the system are readily intelligible to the user. Only if this is true will the user both be able to interact competently and efficiently with the system during its reasoning process, and also be confident in the system's reasoning and advice (Jamar, 2001). And only then can the lawfulness of a decision or recommendation be efficiently or substantively evaluated.

It is now time to bridge the discussion on transparency as an overarching legal value comprising benefits as well as risks to the exercise of fair and lawful public decision-making, over to the utilisation of 'paradata' in this context. Here, the argument is simple enough—that the collection of 'paradata' could, and should, be emphasised as a pro-transparency measure in relation to automated decision-making processes. Data on the data-related processes through which the system works, including on how data are collected and interpreted, is highly relevant for making sure that automated systems produce decisions or recommendations that are in accordance with the law. Analysis of 'paradata' could, for example, help answering important questions like: What processes are in place for the system to retrieve data (including what sources these data are collected from)? What processes does the system use to evaluate whether the collected data are accurate and sufficient to inform a decision, as well as whether further investigation is needed? Are there established feedback mechanisms in place, and how are they designed to work? Is the system equipped with precautionary security measures, such as set procedures for when to interrupt a decision-making process and when it is to be handed over for human review? Did these specific processes run in a particular case of using the system, and how did these different processes combine or feed into each other?

As indicated by the example questions above, 'paradata' does not equal either the data that is fed into the decision-making or decision support system or the system outputs in the form of decisions or recommendations. The collection and analysis of ‘paradata’ alone could therefore not answer, from the perspective of lawfulness, important questions such as if a particular recommendation or decision is legally compliant—unless the data reveals instructions on the system’s running-processes that are contrary to law (such as if the system takes legally irrelevant data into consideration). As ‘paradata’ is data-on-data related processes, its primary function in the context of transparent and scrutable public decision-making is that such data enables the taking of additional factors, other than the current representations of the input data, into consideration. In relation to legal analysis, ‘paradata’ is therefore primarily an auxiliary explanation tool that can help to provide context to analyses of whether there is lawful congruity between a system’s in- and output.

Now, although useful as a tool for analysis, the collecting of ‘paradata’ is not necessarily a straight-forward task. It might be that such data are only readily available in the form of system code, illegible or overwhelmingly technical and detailed to most. From the perspective that transparency is not just about the technical or practical availability of data, one could claim that consideration of the relational aspect of transparency necessitates a certain level of active control of what data are collected (selection) and how it is presented (information design). This makes regulated documentation standards and their design particularly interesting from a public transparency perspective. Not only because such regulated obligations make data retrievable, but also because they provide and give expression to modes of governance on how information on a system’s functioning is to be presented. As we will see in the next section, we can also glimpse a tendency towards specified and increased requirements to document data-on-data related processes.

4 Examples of Legal Requirements on Documenting and Keeping Records on Data-Related Processes

As introduced, ‘paradata’ is neither a legal concept nor a term that is used in regulatory practice. However, a recognition that the collection and review of this type of data can function as a safeguarding measure or tool in relation to automated decision-making processes can be discerned in some regulation.

One example of a regulation containing certain requirements on documenting and keeping records of data-on-data related processes is the EU General Data Protection Regulation (GDPR), which applies to the vast majority of all processing of personal data taking place within the EU (Article 2 GDPR). Personal data is defined as any information relating to an identified or identifiable natural person (Article 4(1) GDPR). The regulation explicitly requires controllers, meaning the natural or legal person which determines the purposes and means of the processing of personal data, to be able to demonstrate how they ensure compliance with the regulation (Article 5(2) GDPR). When personal data are handled via automated processes, this may include documentation on the processes used to ensure that the data are only processed when there is a lawful basis to do so, as well as the keeping of records on how these processes did in fact run in a particular case (to enable *ex post* review of their proper functioning).

Except from a few narrow exceptions, Article 30 GDPR lays down general and explicit requirements to keep and maintain records of any personal data processing activities. These include the keeping of records, for example, on the categories of recipients to whom the personal data have been or will be disclosed to, and a general description of the technical and organisational security measures. Although they include the documentation of some fixed design choices on the processes by which personal data are to be handled, as well as some records of their actual operations—these express requirements are rather limited regarding data-on-data related processes in particular. To demonstrate compliance the responsible controllers may, however, sometimes need or want to document such data irrespective of whether Article 30 GDPR explicitly requires it or not. Demonstrating compliance with requirements such as keeping the personal data accurate and up-to-date may be too complex without the help of different kinds of processing tools—such as data classification tools, data quality tools or data flow mapping tools to determine data lineage, etcetera (Libal, 2021; Wrobel et al., 2017). To opt for documenting the design and use of such processes or tools can therefore help protect controllers in the event of potential violations (Grow, 2018). The GDPR thus both directly and indirectly places obligations on (foremost) controllers of personal data to collect and document some data-on-data related processes and practices.

One potentially even more wide-reaching example of direct regulation prescribing the documentation and keeping of records on data-related processes within the EU is found in the upcoming EU Artificial Intelligence Act (AIA).

This regulation will, notably, only apply to those automated decision-making procedures that run on AI technology. The most extensive documentation requirements will also only pertain to those AI systems considered at risk of having an adverse impact on people’s safety or their fundamental rights (so-called high-risk AI systems). These will most likely be relevant to the bulk of public automated decision-making procedures that are assisted by AI technology, as high-risk systems under the proposal, for example, include any AI system deployed in the areas of access to and enjoyment of essential private services and public services and benefits; law enforcement; migration, asylum and border control management and administration of justice and democratic processes (Article 6 and Annex III AIA).

Any high-risk AI system will be subject to far-reaching technical documentation standards, much focused on the documentation of system processes. A detailed description is not expedient here, but the regulation includes that the provider of a high-risk AI system should provide for documentation of how the AI system will or could interact with external hardware or software, as well as of the system elements and process for its development. The requirements also include the description of the system’s general logic—where key design choices such as the rationale and assumptions made, main classification and optimisation choices as well as the relevance of different parameters (etcetera) are to be documented. The same is true for the system architecture explaining how software components build on or feed into each other and integrate into the overall processing, as well as the computational resources used to develop, train, test, and validate the AI system. The regulation will also require that relevant datasheets describing the training methodologies, techniques, and training data sets used are to be provided. These sheets should include detailed information about the provenance of those data sets, their scope and main characteristics, how they were obtained and selected, any labelling procedures or data cleaning methodologies. Further examples are descriptions of pre-determined changes to the AI system and its performance, detailed information about used validation and testing procedures as well as those relating to monitoring, functioning, and control or risk management. Any changes made to the system through its lifecycle, as well as on the system in place to evaluate the AI system performance are also to be included (Articles 11, 12 and Annex IV AIA).

Notably, all the above-mentioned documentation and recordkeeping requirements pertain to fixed design choices made by the system provider (including by sub-contractors of the provider) that relate to the process operations of the systems. The regulation will, however, also require providers to ensure certain logging capabilities while the system is operating. In contrast to the rather detailed documentation requirements, the required scope and contents of these logging capabilities are not very elaborated. They should, however, ensure a level of traceability of the AI system's functioning throughout its lifecycle (to an extent that is appropriate to the intended purpose of the system). The regulation also states that these logging records should in particular enable the monitoring of certain risks to health or public safety, etcetera, or substantial modifications of the system, and that they should facilitate the 'post-market monitoring' of the system (Article 61 and 3(25) AIA).

As seen, these requirements comprise fixed process design decisions as well as the collection of data on the particular application of these processes, and therefore captures aspects of 'paradata' collection under the definition used in this contribution. One limitation to the regulation is that it is primarily aimed at establishing requirements on providers of AI systems, whereas it is vaguer on extent to which public authorities in the capacity as users, deployers, of high-risk systems is to monitor the systems workings by analysing available data, such as data-on-data related processes. It is clear that system deployers should follow the system instructions and have access to certain information on its functioning (Article 13 AIA). It is also clear that any documentation and recorded data are to be made available to competent supervisory authorities upon request, and that it therefore is meant to facilitate supervisory scrutiny (Article 23 and Recital 46 AIA). In all, it seems that the documentation and recordkeeping requirements of the regulation are primarily geared towards providing for the informational transparency of AI systems, and less around what these data are to be used for and by whom.

It is clear from the upcoming AIA that documentation and the keeping of records—not only on what data these AI systems run on, but also of the data-related processes they premise or perform—has been emphasised. And even if there might remain certain (intentional or inadvertent) gaps in the regulation regarding the utilisation of this data, it is still clear that the overarching aim for the detailed standards, however cumbersome to realise, is to provide an adequate basis for ensuring and monitoring the safe and proper functioning of AI systems.

5 Utilising 'Paradata' for Increased Transparency of Technologically Assisted Public Decision-Making

Automation carries the potential for delivering speedy and more cost-efficient public decision-making but does not reduce the complexity of the law itself. The responsibilities of public authorities to ensure that any decisions they make are in accordance with the law therefore continues to require intermediaries—which now also have to decipher technical information (despite individual differences their aptness to do so) (Čyras & Lachmayer, 2015; Felzmann et al., 2020). At the same time, new technical tools also complement the legal order by offering new means to monitor the side effects of automated decision-making procedures (Fule & Roddick, 2004; Tamò-Larrieux, 2021). Technology may enable the keeping of larger and different sets of records and at lower costs if compared to manual records. While the importance of keeping records in relation to transparent public decision-making procedures is apparent, the problem is rather how to ensure that value is generated through these records (for example, in the form of better reviewability of the public exercise of power). This requires an understanding of what types of analysis the data is meant to support, that there are measures in place to ensure that the relevant data are collected and presented in a way that is legible to human intermediaries.

Different forms of data documentation (electronic or other) have always been imperative to the legal practice. The same is true for data analysis, as the evaluating and assessment of whether a fact—data or sets of data—is relevant, accurate, and substantiated enough to form the basis of a particular decision lies at the core of legal analysis. In addition, this analysis must also capture whether that particular decision came about in a systematic and formal way following certain procedural requirements—ultimately to safeguard the integral structure of the legal system. In relation to both these aspects of legal analysis, the growing use of automated decision-making procedures have somewhat changed the playing field. Where automated processes aim to assist the application of law, the subsumption of legal facts under legal criteria is accomplished by the system (Čyras & Lachmayer, 2014).

The principal merit of analysing 'paradata' in the legal context is that it could help reduce the level of opaqueness and abstraction that these systems display. The interaction taking place between human intermediaries and systems could, however, transpire in relation to different aspects of the system's workings, as well as be performed by different categories of 'humans' with different authorisations as well as knowledge bases. Naturally, the aptness to identify, understand, and make use of relevant data-on-data related processes will also depend on whether the 'human in the loop' is a lawyer, data scientist, or other. The potential to reduce the need for highly qualified personnel, as well the potential to reduce mundane and labour-intensive human administration has been one main reason for the growing deployment of automated systems in performing or assisting public decision-making (Tamò-Larrieux, 2021). Overall, it is therefore unrealistic to expect that everyone involved at all stages of a decision-making process would have the mandate, time, and know-how to utilise any collected 'paradata' to the same degree. The information needed to oversee decision-making processes may hold different degrees of granularity depending on the context, the particular user, and the likely weight of the outcome that the system informs (Oswald, 2018).

On the general level, however, knowledge about the processes in place to evaluate what data a decision is to be based on, or data on the actual process elements that made up the particular procedure by which a decision or recommendation was made may, for example, help the assessment of whether a case has been decided on sufficient grounds. And knowledge on the processes by which the input data has been collected and processed may help the assessment of whether there is reason to question the accuracy of that data in relation to a particular decision. Data on the processes in place to trigger safety-measures such as fall-outs to manual administration, or knowledge on the selection profiles that determines the more specific arrangement of particular process elements, could also help the evaluation of whether the process practices align with procedural requirements and thus the law. 'Paradata' documentation is, thus, one measure to increase the transparency of a system's normative features—improving the reviewability of the process as such, as well as of individual decisions.

Having established that 'paradata' might be useful for legal analysis, this points to the need for making such data readily accessible and useable to human overseers (with different competencies and at different levels of the decision-making procedure). And this is where the design and scope of recordkeeping standards come in. Cobbe argues that the difficulties associated with understanding, overseeing, or reviewing automated decision-making processes often not only suffer from the opaqueness arising from the meeting between technology and people who lack sufficient technical know-how (illiterate opacity), or from the complexity and difficulty of interpreting the system irrespective of technical know-how (inherent opacity). She argues that these systems might also display a type of 'unwitting' opacity stemming from that those responsible for designing, developing, deploying, and using systems simply don't think to record relevant organisational aspects of the system processes (Cobbe et al., 2021). While it should be stressed that the mere recording of different types of process-related data would not help efficiently overcome the opaqueness of automated decision-making procedures, and that 'paradata' documentation requirements certainly is no single or 'silver bullet' solution in this respect—they importantly help to mitigate unwitting opaqueness as a first step towards serving the informational aspect of transparency.

From the perspective of legal analysis, one advantage of documentation standards is that they to some extent require that legal and technical knowledge is meshed and presented in a way that better help the knowledge distribution to users of these systems. Requirements on this type of meshing is seen in the AIA draft requirements, as many of the required entries presuppose the active articulation and augmentation of decision-making processes, rather than the mere disclosure of technical system process data. Although any documentation standards represent a type of selection and prioritisation of certain data or information over other, and although this means that they will create proxies through which a more complex reality of a system's procedure is presented in a comprehensive format—'paradata' documentation and recordkeeping standards is one way to facilitate a common ground for the conversation and conceptions about data-related processes. They may thus also serve the relational aspect of transparency.

6 Conclusions

In the light of all things discussed, it is relevant to question whether 'paradata' as a particular terminology contributes something specific to the legal domain? Here, the strictly formal answer is simple—it does not. The term is not used in any regulation and does not provide any formal or substantive guidance to the content scope of legal recordkeeping standards, either in the GDPR, the upcoming AIA or elsewhere. It is quite possible to analyse data-on-data related processes, as well as to set requirements which include keeping records on data-related processes, without specifically framing this as 'paradata' analysis or 'paradata' collection.

From the perspective of legal analysis, however, 'paradata' as terminology could serve a pedagogical function in distinguishing different types of data from each other and aid better cognisance of what types of analysis on decision-making procedures that it may support. 'Paradata' is not the input data that is used to feed an automated decision-making procedure. 'Paradata' is also not that type of data that describes and gives information about other data, such as information on when and by whom a certain data was collected (metadata). 'Paradata' is data-on-data related *processes*. It may therefore provide information on the procedural aspects of automated decision-making processes (as particularly relevant in relation to the public exercise of power).

The here used definition of 'paradata' is intentionally broad and includes data on fixed design decisions on system processes, as well as data on how these design decisions have been applied in particular applications (that is, in individual decisions). There are overlaps with similar terminology such as contextual metadata, or statistical and process data, in that they include process-related data. More important than the specific definition or choice of terminology is, however, to point to the functional need to collect and analyse data-on-data related processes. The context-creating merit of attributing certain types of data specifically to 'paradata' lies, at least from the legal analytical point of view, in that it conflates data with procedural properties into a cohesive category of information—around which awareness and knowledge on data-related processes could be more effectively managed.

So, although 'paradata' is neither a fixed term nor a fixed legal concept—and irrespective of whether it will ever permeate into the legal vocabulary—it has a clear utility function in relation to legal knowledge management and the data analysis imperative to ensure that automated decisions or recommendations align with the law. It is evident that the mere keeping of records that include 'paradata' does not solve the problems of opaque decision-making procedures. The clear challenges to utilising 'paradata' in legal analysis are distinct and undeniable. These include competence issues (the technical knowledge to decipher the data and relate this information to legal requirements). Challenges also include the organisational conditions within the public authorities regarding whom and how oversight is to be performed. Some of these challenges could be addressed by legislative measures that do not relate to documentation or the keeping records. Records are, however, still at the core of the legal infrastructure, and perhaps even more so in the age of technology. As put by Iacovino, recordkeeping lies at the heart of some of the fundamental assumptions of how and why legal systems develop and is not only supported by—but also supports—the practice of the law (Iacovino, 1998). The establishment of recordkeeping regimes that are able to assist the structural and substantive qualities of legal system is therefore a topic deserving of much more in-depth attention in the age of accelerated technological assistance in public decision-making.

Acknowledgements This work was supported by the Swedish Research Council under Grant number 2020-02278.

References

AIA. At the time of writing, the final text of the AIA has not yet been adopted. References to the AIA in this chapter are based on the February 2024 text of the provisional agreement resulting from interinstitutional negotiations between the European Parliament and the EU Council of Ministers. This text outlines the content of the Regulation but may undergo minor, primarily editorial changes before final adoption. Accessed May 15, 2025, from https://www.europarl.europa.eu/meetdocs/2014_2019/ plmrep/COMMITTEES/CJ40/AG/2024/02-13/1296003EN.pdf.

Bentkowska-Kafel, A., Denard, H., & Baker, D. (2012). *Paradata and transparency in virtual heritage*. Ashgate.

Berring, R. C. (1994). Collapse of the structure of the legal research universe: The imperative of digital information. *Washington Law Review, 69*(1), 9–34.

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data and Society, 3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

Carney, T. (2019). Robo-debt illegality: The seven veils of failed guarantees of the rule of law? *Alternative Law Journal, 44*(1), 4–10. https://doi.org/10.1177/1037969X18815913

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). *Reviewable automated decision-making*. ACM. doi:https://doi.org/10.1145/3442188.3445921.

Couper, M. P. (2017). Birth and diffusion of the concept of paradata. *Advances in Social Research, 2017*(18), 14–26.

Čyras, V., & Lachmayer, F. (2014). Program transparency for legal machines. In *Jusletter IT* (Vol. 2014, pp. 47–54). Presented at the *Transparency, Proceedings of the 17th international legal informatics symposium, IRIS 2014*, 20–22 February 2014, Universität Salzburg, Wein: Österreichische Computer Gesellschaft. https://doi.org/10.38023/2d0f8625-a63b-411d-93d9-f1262a443811.

Čyras, V., & Lachmayer, F. (2015). Towards multidimensional rule visualizations. In M. Araszkiewicz, P. Banaś, T. Gizbert-Studnicki, & K. Płeszka (Eds.), *Problems of normativity, rules and rule-following* (pp. 445–455). Springer International Publishing. https://doi.org/10.1007/978-3-319-09375-8_33

Davis, A. E. (2020). The future of law firms (and lawyers) in the age of artificial intelligence. *Revista Direito GV, 16*(1). https://doi.org/10.1590/2317-6172201945

Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review, 119*(7), 1829–1850.

du Plessis, T., & du Toit, A. S. A. (2006). Knowledge management and legal practice. *International Journal of Information Management, 26*(5), 360–371. https://doi.org/10.1016/j.ijinfomgt.2006.06.003

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law and Technology Review, 16*(1), 18.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamo-Larrieux, A. (2019a). Robots and transparency: The multiple dimensions of transparency in the context of robot technologies. *IEEE Robotics & Automation Magazine, 26*(2), 71–78. Presented at the IEEE Robotics & Automation Magazine. doi:https://doi.org/10.1109/MRA.2019.2904644.

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019b). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 6*(1), 205395171986054. https://doi.org/10.1177/2053951719860542

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics, 26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Foster, L., & Kennedy, B. (2000). Technological developments in legal research the evolution of research. *Journal of Appellate Practice and Process, 2*(2), 275–304.

Fule, P., & Roddick, J. F. (2004). Detecting privacy and ethical sensitivity in data mining results. In *Proceedings of the 27th Australasian conference on Computer science* (Vol. 26, pp. 159–166). Australian Computer Society,. Accessed 14 September 2022.

Grow, G. (2018, March 28). *Understanding the role of data quality in GDPR article 5 compliance. Best Data Management Software, Vendors and Data Science Platforms*. Accessed September 14, 2022, from https://solutionsreview.com/data-management/understanding-the-role-of-data-quality-in-gdpr-article-5-compliance/

Iacovino, L. (1998). The nature of the nexus between recordkeeping and the law. *Archives and Manuscripts, 26*(2), 216–246. https://doi.org/10.3316/informit.294192853020461

Independent High-Level Expert Group on Artificial Intelligence (Trans.). (2018). *Ethics Guidelines for Trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Jamar, S. D. (2001). The human right of access to legal information: Using technology to advance transparency and the rule of law. *Global Jurist, 1*(2), 6. https://doi.org/10.2202/1535-167X.1032

Kidd, A. (1985). Chapter 14 - Human factors problems in the design and use of expert systems. In A. Monk (Ed.), *Fundamentals of human–computer interaction* (pp. 237–247). Elsevier. https://doi.org/10.1016/B978-0-12-504582-7.50023-8

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review, 9*(2), 1–16. https://doi.org/10.14763/2020.2.1469

Libal, T. (2021). Towards automated GDPR compliance checking. In *Trustworthy AI - Integrating learning, optimization and reasoning* (pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-73959-1_1

Matheus, R., Janssen, M., & Janowski, T. (2021). Design principles for creating digital transparency in government. *Government Information Quarterly, 38*(1), 101550. https://doi.org/10.1016/j.giq.2020.101550

Merwe, D. P. V. D. (1986). *Computers and the Law*. Juta.

Mock, W. (1999). On the centrality of information law: A rational choice discussion of information law and transparency. *The John Marshall Journal of Computer and Information Law, 17*(4), 1069–1100.

Ng, Y.-F., O'Sullivan, M., Paterson, M., & Witzleb, N. (2020). Revitalising public law in a technological era: Rights, transparency and administrative justice. *University of New South Wales Law Journal, 43*(3), 1041–1077. https://doi.org/10.53637/YGTS5583

Oswald, M. (2018). Algorithm-assisted decision-making in the public sector: Framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of* *the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences, 376*(2128), 20170359. https://doi.org/10.1098/rsta.2017.0359

Reilly, P., Callery, S., Dawson, I., & Gant, S. (2021). Provenance illusions and elusive paradata: When archaeology and art/archaeological practice meets the phygital. *Open Archaeology (Berlin, Germany), 7*(1), 454–481. https://doi.org/10.1515/opar-2020-0143

Roos, J., Roos, G., Dragonetti, N. C., & Edvinsson, L. (1997). *Intellectual capital: Navigating the new business landscape*. Macmillan.

Susskind, R. E. (2000). *Transforming the law: Essays on technology, justice and the legal marketplace*. University Press.

Susskind, R., & Susskind, D. (2016, October 11). Technology will replace many doctors, lawyers, and other professionals. *Harvard Business Review*. Accessed September 19, 2022, from https://hbr.org/2016/10/robots-will-replace-doctors-lawyers-and-other-professionals

Tamò-Larrieux, A. (2021). Decision-making by machines: Is the 'Law of Everything' enough? *The Computer Law and Security Report, 41*, 105541. https://doi.org/10.1016/j.clsr.2021.105541

Wrobel, A., Komnata, K., & Rudek, K. (2017). *IBM data governance solutions* (pp. 1–3). IEEE. https://doi.org/10.1109/BESC.2017.8256387

Zalnieriute, M., Moses, L. B., & Williams, G. (2019). The rule of law and automation of government decision-making. *The Modern Law Review, 82*(3), 425–455. https://doi.org/10.1111/1468-2230.12412

Lena Enqvist is an Assistant Professor of Law at the Department of Law, Umeå University. She is currently engaged in research related to digitalisation and automation in the public sector. This research encompasses inquiries into the necessity, function, and conditions of legal frameworks in fostering transparency and enabling accountable exercise of authority.

Concluding Discussion: Paradata for Information and Knowledge Management

Isto Huvila, Lisa Andersson, and Olle Sköld

Abstract

This concluding chapter draws together insights from the discipline-specific chapters to contrast and synthesise the diverse approaches to how the concept of paradata is conceptualised and used in the different cases covered in the volume "Perspectives to paradata". Paradata as a concept that refers to process information resides firmly at the fringe of codified knowledge and organisational learning. Many different forms of information can function as paradata. There is a comparable variety in how the concept of paradata is understood. Due to the variety, having a unified definition can be debated. Major opportunities with paradata range from achieving reproducibility in data analysis and use and delivering the desired outcomes of the Open movement to increasing algorithmic and administrative accountability and transparency of artificial intelligence. However, transparency brought by paradata is not automatically virtuous. Paradata ethics and its relation to general information and knowledge management ethics is central to responsible use of paradata. Besides further inquiry into paradata concept, actual instances of paradata in the wild and how they are linked to social action, it is a key area that requires further research.

I. Huvila (✉) · L. Andersson · O. Sköld
Uppsala University, Uppsala, Sweden
e-mail: isto.huvila@abm.uu.se; lisa.andersson@ki.se; olle.skold@abm.uu.se

I. Huvila et al. (eds.), *Perspectives on Paradata*,
Knowledge Management and Organizational Learning 13,
https://doi.org/10.1007/978-3-031-53946-6_14

1 Introduction

Process knowledge has many faces. Its diversity parallels with and simultaneously goes far beyond the multiplicity of processes and practices themselves. In this volume, our aim has been to delve into this diversity and its implications by introducing and exploring the notion of paradata as a concept and practical tool in and for managing knowledge—whether it is a matter of transferring knowledge between two archaeologists standing next to each other in the field, promoting societal accountability, or recording a computational research code to enable reproducibility of scientific discoveries. Doing so, we expand the insight into how the paradata concept is understood in different disciplines and cases, and why paradata have proven useful in practice to answer questions or achieve goals with data, information, and knowledge. By this exploration of process knowledge, how process information can support the management of information and knowledge in different forms, and of how process information and knowledge can be managed in a variety of contexts we hope to contribute to theoretical and practical advancement in the field of information and knowledge management.

The work with this volume commenced with a working definition of paradata that all the chapter authors were asked to reflect but not necessarily agree upon. It has served as a common ground and a point of departure to discipline- and context-specific explorations of what paradata can be in different settings, what the character of the processes is meant to describe, what methods are used to find or generate paradata, what paradata can do or enable, and what needs to be considered when creating and using paradata in different contexts.

This concluding chapter draws together insights from the discipline-specific chapters to contrast and synthesise the diverse approaches to how the concept of paradata is conceptualised and used in the different cases covered in this volume. Further, we proceed to three topics of discussion emerging from the synthesising analysis. First, we discuss how paradata are done in practice by various actors using different paradata creation methods for their specific purposes. Second, we delve into the implications of paradata for the theory and practice of information and knowledge management. Before concluding with brief remarks on future directions of paradata research and practice, this chapter spends a few words to discuss a third and crucial question that remains somewhat implicit in the chapters. It is the one of ethics of paradata and potential ethical hurdles that need to be considered when paradata—descriptions of activities—are put into practice.

2 Paradata: In Plural

While the punchline of this volume is that paradata, a potentially useful concept, has been under-researched so far, we discussed already in the introduction how the notion is by no means new. Chapter “Paradata in Surveys” (Schenk and Reuß) and chapter “A Leap of Faith: Revisiting Paradata in 3D Scholarship” (Papadopoulos) elucidate the history of the paradata term in two disciplines where it has the longest history of use: survey research and 3D heritage visualisations. While it is impossible to pinpoint the exact reason why paradata emerged first in survey research and sometime later in heritage contexts, it is reasonable to state that it has been useful in relation to particular kinds of documentation needs evident in these two contexts. Survey scholars have needed to account for the processes going into a certain survey dataset, and heritage scholars making 3D reconstructions have needed to communicate the technical details and decisions going into creating a 3D model. The chapters of this volume show that related needs can be identified also in other contexts, although similarly to how both the understanding of what paradata are and how they are supposed to be acted upon differs between survey research and heritage studies, the desideratum relating to paradata and its practical role differs between particular contexts.

Some of the chapters note that the term paradata has been established in the disciplinary discourse of their respective domains. This applies to the above-mentioned pioneering fields of survey research and heritage visualisation but increasingly also to research data management and information research. In others, paradata as a concept is not a part of the conceptual apparatus of the discipline or has only recently been introduced. For example, Enqvist remarks forthright in chapter "Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes" that in the legal domain, the paradata concept does not exist. When paradata is not a part of the formal requirements and vocabulary of particular fields, it is best described as a complementary lens to understanding specific types of information or data and their function (e.g., chapters "Making Research Code Useful Paradata", "The Role of Paradata in Algorithmic Accountability", "Adding Paradata About Records Processes via Information Control Plans", and "Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes"). In those cases, introducing paradata as a conceptual tool can be a way to delineate particular forms of knowledge that has not yet been systematically recorded (as for process information about computational code, AI algorithms and automated decision-making).

Another indication of the varying status and genealogy of paradata are the different origin stories of paradata outlined in the different chapters, most explicitly in chapters "Paradata in Surveys", "A Leap of Faith: Revisiting Paradata in 3D Scholarship", and "Mapping Accessions to Repositories Data: A Case Study in Paradata", and how they relate to earlier conceptual overviews in the literature (e.g., Denard, 2012; Edwards et al., 2017; Huvila, 2022; Cameron et al., 2023). While these accounts acknowledge cross-disciplinary influences to a varying degree, a common trait in the chapters in this volume and the earlier literature alike is that the common ground is still fairly weak and the concept is made meaningful through domain-specific appropriations rather than interdisciplinary consensus on the nature of the concept. Independent of the ultimate necessity or desirability of such a consensus, one of the reasons for the heterogeneity is undoubtedly the fact pointed out by Dawson and Reilly in chapter "Towards Embodied Paradata. A Diffractive Art/Archaeology Approach" that paradata have been theorised fairly little so far. This has led to a contrast between distinctly practical and theoretical takes on paradata—for instance in how paradata are discussed largely as a hands-on matter of how to document development of datasets in survey research versus how it is approached by Dawson and Reilly as a manifestly theoretical concept. A parallel distinction can be sensed between a strive for general and definitive resource descriptions, for instance, in research data, archives and records management, and paradata created to situated data reuse needs in heritage visualisation and computational research.

Considering the aims of this volume to illustrate and elicit diversity, it is perhaps unsurprising that paradata appears as a rather amoeba-like concept both in theoretical and practical sense. Paradata, as data, is clearly a plural rather than a singular entity both literally and in what it can be and do in different settings. This does not mean, however, that the different threads do not entail related elements and starting points to a common exploration of what paradata as a whole can imply for information and knowledge management.

3 Doing Paradata

In the chapters much of the described documenting of processes is really about document*ing* rather than documents and repurposing information for new insights rather than generating entirely new data. Therefore, to untangle the knot of how it makes sense as a management concept it is appropriate to start by looking into how paradata are generated and how they come into being in different contexts. Similarly to how paradata are conceptualised in the texts and what is referred to as documentation, the chapters unfold a diverse array of methods on how paradata are or could be achieved in different contexts. Many of the methods overlap. Data that eventually can be useful as paradata may be a by-product of an administrative or a scholarly practice but before the data can function as paradata, they sometimes need to be extracted or (re)purposed as such.

In some of the chapters, paradata are *intentionally generated* through using specific methods for process description. Dillen (chapter “Paradata for Digitization Processes and Digital Scholarly Editions”) and Schenk and Reuß (chapter “Paradata in Surveys”) discuss several explicit approaches of paradata generation from taking notes and collecting ratings to providing a description of editorial principles in a digital scholarly edition. A common observation in chapters and contexts throughout this volume is that while purposeful generation of meaningful and useful paradata is often necessary, it is both difficult and resource-intensive to generate. As a partial remedy, multiple chapters, including those of Dillen (chapter “Paradata for Digitization Processes and Digital Scholarly Editions”), Rayburn and Thomer (chapter “Reconstructing Provenance in Long-Lived Data Systems: The Challenge of Paradata Capture in Memory Institution Collection Databases”), Jones and Bunn (chapter “Mapping Accessions to Repositories Data: A Case Study in Paradata”), and Cohen and colleagues (chapter “Paradata in Emergency Services

Communications Systems") describe how paradata can be *extracted*—or perhaps, more correctly constructed—*post hoc* of the available data and documentation.

In other cases, paradata can *result* from data being collected using particular tools and methods. As for example, the chapters of Schenk and Reuß (chapter "Paradata in Surveys"), Dawson and Reilly (chapter "Towards Embodied Paradata. A Diffractive Art/Archaeology Approach"), Papadopoulos (chapter "A Leap of Faith: Revisiting Paradata in 3D Scholarship"), and Dillen (chapter "Paradata for Digitization Processes and Digital Scholarly Editions") evince, datasets, images, and visualisations all tend to contain plentiful traces of how they were made. Paradata can also be a *by-product* of a process or practice. This is apparent not least in the work of Bilderbeek (chapter "Making Research Code Useful Paradata") where computer code is to a certain extent both practice and documentation. Jones and Bunn (chapter "Mapping Accessions to Repositories Data: A Case Study in Paradata") show further how paradata can be a by-product also in the sense that the original purpose of generating such data was completely different. Packalén and Henttonen's (chapter "Adding Paradata About Records Processes via Information Control Plans") information control plans are another example of how paradata can be a by-product of process planning rather than an independent task of its own. Finally, with a conceptually different approach to paradata, also Buchanan and Huntsman (chapter "Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites") describe by referring to "dustings" something that can be linked to extracting paradata.

The different mechanisms explored throughout the chapters of how paradata either potentially or actually can come into being have obvious affinities to previous categorisations. The earlier literature distinguishes between *ex ante* and *post hoc* (forensic) means of acquiring paradata as well as automatic and manual methods (Huvila, 2022). In fields where paradata—or process documentation in general—are recognised concepts and practices, there are established arrays of specific methods to work with them. In this volume, Schenk and Reuß (chapter "Paradata in Surveys") provide a long list of typical techniques for capturing paradata in survey research including capturing time stamps, keeping call records, location and device paradata, tracking inputs and collecting ratings and observations. Also, here it is possible to see a distinction between purposive paradata generation and collecting by-products of especially digitally administered surveys. In heritage visualisation, where paradata has been discussed since the turn of the millennium, the array of preferred methods is still unfolding as Papadopoulos notes in chapter "A Leap of Faith: Revisiting Paradata in 3D Scholarship" and Dawson and Reilly in Chapter "Towards Embodied Paradata. A Diffractive Art/Archaeology Approach". Disciplines where paradata and process documentation do not have a self-evident role, it is unsurprising that also the methodological apparatus remains less systematic and developed. The flipside of lack of established standards for paradata documentation is, as for example illustrated Rayburn and Thomer in Chapter "Reconstructing Provenance in Long-Lived Data Systems: The Challenge of Paradata Capture in Memory Institution Collection Databases", the leeway to experiment with several different ways of documenting and visualising data processes.

The present volume contains several examples of the split between ex ante and post hoc methods. Trace and Hodges (chapter "The Role of Paradata in Algorithmic Accountability") distinguish, on the one hand, different records created before, during, and after an AI system is deployed, and forms of paradata such as Explainability Fact Sheets and Data Statements to inform, among others, ex ante their designers and post hoc their users. Rayburn and Thomer (chapter "Reconstructing Provenance in Long-Lived Data Systems: The Challenge of Paradata Capture in Memory Institution Collection Databases") show how careful post hoc analysis can provide a lot of insights in how a database was created and how it has been used. On the other hand, Packalén and Henttonen (chapter "Adding Paradata About Records Processes via Information Control Plans") show how an ex ante planning and documentation of processes results in actionable paradata. Besides technical differences, they differ in a fundamental manner in what type of paradata they generate. Narrative descriptions of past activities result in different kind of paradata than a plan of action, or real-time timestamps and coordinates collected in the heat of action.

The chapters contain similarly a rich array of examples of the automatic and manual methods and approaches that are probably best described as hybrids. The examples of documents in the list of Trace and Hodges (chapter "The Role of Paradata in Algorithmic Accountability") on documents that provide paradata for understandability and explainability demonstrate how a specific form of document can function as paradata both to describe previous undertakings and inform future actions. Similarly, even if plans—discussed both by Trace and Hodges (chapter "The Role of Paradata in Algorithmic Accountability") and Packalén and Henttonen (chapter "Adding Paradata About Records Processes via Information Control Plans")—constitute at face value a source of an ex ante form of paradata, depending on when they are finalised, they unfold as descriptions of how processes were planned to be. Buchanan and Huntsman's (chapter "Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites") concept of dustings as narrations of paradata provides a parallel perspective to the temporality and hybridity of paradata by underlining the different timelines of when practices and paradata take place independently of whether it is documented before, during, or after—or as a hybrid—before, during, *and* after the data happen.

Again similarly to how specific approaches operate on different levels of temporality, the different types of methods generate different types of paradata. In parallel, it is also evident how the means of making paradata happen link to diverse forms of information work discussed in the chapters. Where automatic paradata generation is often a straightforward process of collecting and ingesting paradata, it is evident in manual and hybrid processes how they are not only about stockpiling information but involve, incorporate, and prompt reflection and co-shape the data-making. Both Dawson and Reilly's (chapter "Towards Embodied Paradata. A Diffractive Art/Archaeology Approach") and Buchanan and Huntsman's (chapter "Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites") chapters show how this co-shaping can benefit from multidisciplinary perspectives and involvement in the process. At the same time, it is evident from the chapters how involving an artist or a data archivist leads to very different types of observations and paradata. It differs by the techniques of how paradata are created, what types of artefacts are considered paradata, and on a fundamental epistemic level, how paradata come into being.

Finally, as with definitions of paradata, specificity undeniably makes both paradata and the practices of making paradata easier recognisable. Having a named and known technique for collecting paradata like with Data Statements (Trace and Hodges, chapter "The Role of Paradata in Algorithmic Accountability") or EER diagrams (Rayburn and Thomer, chapter "Reconstructing Provenance in Long-Lived Data Systems: The Challenge of Paradata Capture in Memory Institution Collection Databases") contributes to the clarity of what is supposed to be collected and achieved. With some caution, also the digitality of generated paradata can help to contribute to the transparency of paradata generation and the findability of resulting paradata. A possible downside is how they might narrow down the understanding of what might count as paradata much similarly to how paradata have so far often been neglected in formal documentation. Less specific approaches such as encouraging individuals to narrate their doings or exploring artistic methods can sometimes be better in capturing the fluidity of processes and practices but at the same time, increase the variegation of how paradata are done and what the resulting paradata are and are capable of achieving.

An attempt to summarise the variety of approaches to doing paradata in this volume is obviously difficult. A glimpse to the means of how paradata are often literally done in practice—even if it is also sometimes explicitly, for example, collected, generated, or made—shows how understanding what paradata entail in diverse situations and contexts is not only dependent on how it is conceptualised. It is also as much dependent on how it is made and acted upon in practice. We posit that these two viewpoints, or perhaps rather constellations of vistas, are also a key in opening up perspectives to not only actual paradata but also where paradata as a concept and an arrangement of practices places itself in the context of information and knowledge management.

4 Paradata for Information and Knowledge Management

Considering the theorising and practical use cases for paradata explored in this volume, we propose that there is place for paradata in the conceptual apparatus of information and knowledge management. Depending on how it is understood conceptually and operationalised in practice, it can be fitted in the major discourses of the field as a tacit understanding of processes and practices turned to a data-thing that is manageable in a knowledge management system, or used as a concept that stands for the tacit understanding that needs to be managed through a social and socio-technical mesh of people and technologies (cf. Handzic, 2004). Before rushing into conclusions of how and where paradata might be placed in relation to the canon of information and knowledge management terminology, it is useful to step back and consider what the apparent openness and pliability of the concept might imply for paradata in relation to managing of both information and knowledge.

We readily acknowledge that information termed in this volume as paradata can often be described using neighbouring concepts, especially if the scope of inquiry is limited to a single discipline, like archival science in chapters "Mapping accessions to repositories data: A case study in paradata" and "Adding paradata about records processes via Information Control Plans", or archaeology in chapters "Reconstructing provenance in long-lived data systems: the challenge of paradata capture in memory institution collection databases" and "Towards Embodied Paradata. A diffractive art/archaeology approach". Similarly, again in specific disciplinary contexts it is possible to see apparent overlap between what can be conceptualised and treated as paradata and what has been traditionally termed as something else. However, as Dawson and Reilly suggest in chapter "Towards Embodied Paradata. A Diffractive Art/Archaeology Approach" by discussing the notion of peridata, it is possible that even paradata might not be enough to address the complexities of documenting processes and practices in detail.

Even if there should be no rush to abandon earlier conceptualisations or perspectives to process knowledge, information, and data, the chapters in this volume point to a plethora of advantages of working *with* the notion of paradata. Most importantly, it can help to bring forth and make explicit aspects of processes that can be difficult to recognise, frame, and discuss when they are treated as a part of something else whether it would be a general description, context, or the historical origins of the object of interest. As Enqvist (chapter "Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes") remarks, while paradata is not a legal term and thus incapable of providing any formal guidance in the legal domain, it can still serve a pedagogical function. Another example is how Trace and Hodges (chapter "The Role of Paradata in Algorithmic Accountability") use paradata as a lens to understand how accountability is conceptualised in relation to algorithmic systems, and a third one, how Jones and Bunn (chapter "Mapping Accessions to Repositories Data: A Case Study in Paradata") have used the notion to inquire into a dataset and repurpose to a new use.

In this edited volume we have intentionally widened the range of disciplines in which the paradata concept is applied and approached, not only as a matter of information and knowledge management for the information and knowledge management field, but for the management of information and knowledge in and across disciplinary contexts far beyond that. Some of the disciplinary contexts feel perhaps more "natural" in this respect as the reflections on the novelty and currently more or less established status of the notion in different disciplines evince. Others required more explicit nudging by us and the chapter authors to open up for testing the term and its usefulness for the purpose of conceptual exploration of the discipline-specific practice. Similarly, in some of the chapters the links to information and knowledge management—in a broad sense including management of data, records, and other informational assets, processes, practices, and doings—are more explicit whereas in others, they might remain rather implicit. However, it is also equally evident that paradata are always to a certain extent related to management of something independent of the disciplinary setting, like the management of a dataset (chapter “Making Research Code Useful Paradata”), a database (chapter “Reconstructing Provenance in Long-Lived Data Systems: The Challenge of Paradata Capture in Memory Institution Collection Databases”), or digitised texts (chapter “Paradata for Digitization Processes and Digital Scholarly Editions”).

Whether explicitly referring to information and knowledge management or not, the chapters make multiple references to diverse examples of what paradata can make manageable, what can be managed with paradata, and how paradata themselves can be managed. Paradata can be described as a management concept comparable to others that function as a prism in surfacing and aid thinking about process descriptions, highlighting many otherwise invisible or forgotten facets of cognition, interactions, negotiations, intangible, embodied, unconscious, unregarded, or blinded processes across contexts from archaeology (e.g., chapters “Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites” and “Towards Embodied Paradata. A Diffractive Art/Archaeology Approach”) to artificial intelligence (chapters “Mapping Accessions to Repositories Data: A Case Study in Paradata” and “The Role of Paradata in Algorithmic Accountability”). It can similarly be conceptualised as and compared to a boundary object (chapter “Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites”), friction point (chapter “Making Research Code Useful Paradata”) and as throughout the volume, an interface between practices, processes, and their related entities. Perhaps in the least formal sense, paradata can serve as a tool for self-reflection, reflecting upon reliability, validity, reproducibility of processes and practices across contexts.

The chapters that take an empathetically technical rather than theoretical take on paradata in highlighting how it can matter in practice. Chapter “Paradata for Digitization Processes and Digital Scholarly Editions” shows how paradata can be helpful managing library collections when they are being digitised. Like paradata can in conceptual sense make processes thinkable, it can help to make putting processes into words, accessible, analysable, and changeable. Chapters “Paradata in Emergency Services Communications Systems” and “Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes” underline how paradata are intimately linked to information and control in the context of legal processes and emergency services, and chapter “Towards Embodied Paradata. A Diffractive Art/Archaeology Approach” how paradata can systematise the understanding of the temporality and ephemerality of processes beyond stating the fact. As, for example, the chapters of Bilderbeek (chapter “Making Research Code Useful Paradata”) and Jones and Bunn (chapter “Mapping Accessions to Repositories Data: A Case Study in Paradata”) evince how paradata can function as a measure against the obsolescence of information and knowledge by providing means to understand it and sometimes, in practice, (re)create it, if necessary, from new premises.

***

What then makes paradata pertinent for information and knowledge management right now? The ubiquity of large-scale data processing in broad areas of social life has increased the calls for the importance of reproducibility of analyses that underpin decisions and knowledge-making. Both Trace and Hodges (chapter "The Role of Paradata in Algorithmic Accountability") and Bilderbeek (chapter "Making Research Code Useful Paradata") highlight the opportunities with paradata in achieving reproducibility.

Reproducibility ties also to the broader issue of the new advent of Artificial Intelligence that has made paradata perhaps more pertinent than ever. Blackboxing has become increasingly apparent in society and linked to an unprecedented array of modes of social action. An advantage of the paradata concept has grown and is growing especially in contexts where blackboxing has been experienced as a problem already before. While much more than mere paradata are needed to address the conundrums of the transparency of algorithms, thinking with paradata can help to work towards a greater transparency of artificial intelligence techniques and algorithms (Trace and Hodges, chapter "The Role of Paradata in Algorithmic Accountability" and Bilderbeek, chapter "Making Research Code Useful Paradata"; cf. Cameron et al., 2023). In a wider sense, from the perspective of using paradata as a lens to documentation and management of process information and knowledge, it can help to take necessary steps to shed light at least to parts of black boxes without losing the sight of the complexity of what it takes to make a process transparent enough to be intelligible.

The opposite face of the pertinence of paradata in the contemporary information and knowledge landscape is how it can also help to distinguish between the knowledge about processes someone wants to share and withhold. Enqvist's chapter "Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes" draws attention to how legislation affects and reflects the strive to open and close. Outside of the scope of the themes touched upon in this volume, for example patents and lab notebooks provide further examples of settings and techniques of how to regulate transparency with something that could be termed paradata.

***

While we acknowledge that we have but touched the possible settings and situations where paradata might matter, as we suggested already in the introduction, we posit that in the context of information and knowledge management the paradata as a concept resides firmly at the fringe of codified knowledge and organisational learning. Paradata makes sense both as a lens to make visible the complexity of processes and practices, and the limits to what extent they can be codified, and as a form of documentation that catches the complexity, systematicises it, and makes it intelligible across time, as is illustrated, for example, by the paradata for database maintenance in chapter "Reconstructing Provenance in Long-Lived Data Systems: The Challenge of Paradata Capture in Memory Institution Collection Databases". When successfully implemented, paradata-as-codified-knowledge (or paradata-as-data) unfolds as an asset for pushing forward the ideals of Open Science, friction-free data publishing and sharing, and systematic data governance.

At the same time, a closer look at paradata-as-lens underlines the limits of such attempts and directs attention to the situated nature of processes, practices, and how they can be knowable to anyone not part of them.

While these two standpoints might appear irreconcilable, they do also represent two perspectives to how and to what extent paradata are difficult and possible to achieve. In this sense, paradata is clearly a disruptive concept as it both suggests and resists the idea that data processing information can be codified, captured, and passed along on a time-space continuum. Both perspectives make paradata transformative in their respective manners. Chapter "Mapping Accessions to Repositories Data: A Case Study in Paradata" highlights how paradata-as-data has a power to datafy, or to turn data to data in a particular sense whereas, perhaps especially, the work of Dawson and Reilly in chapter "Towards Embodied Paradata. A Diffractive Art/Archaeology Approach" points to how considering the conceptual premises and implications of paradata points to the opposite. However, as Stephanie Bunn et al. (2022) notes, while it is tempting to juxtapose "the kinds of knowledge we think can be extracted or condensed from a craft process and transferred into a diagram for a novice learner" and the ones impossible, it is necessarily not the most productive approach. The chapters show how in line with the technical perspective to managing knowledge, paradata unfold as a potential tool helping to make tacit knowledge explicit (cf. Polanyi, 2009; Nonaka & Takeuchi, 1995), and from a human perspective, paradata are something to make tacit knowledge, that is fundamentally non-codable, easier to understand, and following the classification of knowns and unknowns of Huggett (2020), to make unknown unknowns at least known unknowns. The chapters themselves feature multiple examples how the two perspectives can be combined in a reflexive dialogue, for example, in digital textual and visual scholarly editions. The bottom line is not paradata itself, but about doing things with data, critical reflection on such doings and their implications, and making them understandable to an extent that is deemed desirable.

5 The Idea of Transparency and Ethics of Paradata

When considering the implications and opportunities of paradata, it is necessary to direct attention to an aspect of paradata that has remained largely implicit throughout much of this volume. This is the extent to which processes and practices are indeed desirable to be made transparent and what ethical concerns paradata arises in par with trying to solve others. While paradata is admittedly a pro-transparency concept and intuitively about (positive) openness, increased trust (e.g., chapters "Dustings of Paradata as Pedagogical Support at Four Archaeological Field-School Sites" and "Mapping Accessions to Repositories Data: A Case Study in Paradata"), and accountability (chapter "Adding Paradata About Records Processes via Information Control Plans"), there is nothing in any form or notion of paradata that makes it automatically virtuous. Even if it would be tempting to rally for what Hess (2005) describes as a technology- or product-oriented movement to promote paradata as a definite means for positive change, there is reason for caution. There can be both too much and in different terms “wrong” kind of paradata that is inappropriate, or for example, difficult to navigate for their users. Enqvist’s discussion in chapter “Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes” raises a legitimate and highly pertinent question of whether transparency is always desirable. It is necessary to hold back and consider for a moment what eventually makes paradata ethical and responsible.

Similarly to how paradata is conceptually multi-faceted and knotty, also its ethical underpinnings and repercussions are diverse. An obvious dividing line is how and to what extent paradata is approached as a lens or a form of data. They can be roughly seen following what Mickel and colleagues (2023) distinguish as two cultures of ethics “one focused on the relationship between science and social justice, and the other focused on clean data and accuracy as an ethical issue”. However, as they continue, also the mundane interactions shape “at a decidedly local level” how those who engage in everyday work “understand and ultimately approach ethical decision-making”. They, for their part, shape everyday practices with paradata that shape paradata itself and what is considered as good practice.

In addition, such issues as the availability and unavailability of paradata have implications to how (para)data is understood, what kind of new (para)data is generated, what it makes manageable and how, and what consequences it has on different groups and individuals. The reasons to create and eventually (not) share data have also implications to paradata, what they imply and to whom. Legal professionals have theirs, archivists theirs, and researchers their multiple context and situation influenced reasons to embrace the concept and particular practices to operationalise it. For researchers, data sharing for the purposes of gaining credit or economic benefits has very particular implications to what paradata end up being and doing with an emphasis on the gain rather than documentation. Altruistic sharing and documentation motivated by helping others by sharing an existing resource has others—perhaps focused on very particular ideas of transparency— similarly to obligatory creating and sharing that might lead to prioritising conformity with guidelines rather than usefulness and responsibility.

There is a plethora of conceivable ethical risks with both poor and too specific paradata. While paradata can counter obsolescence also it can be and become obsolescent and indecipherable. Paradata is easy to use for measurement, assessment, and evaluation of practices and processes beyond what is fair and reasonable. Producing paradata can be used as a token for underlining general commitment to transparency even if the documentation itself would remain of limited quality. Paradata comes similarly with a risk of misleading people by emphasising aspects of processes and data that might not be the most pertinent ones. Paradata come with an imminent risk of surfacing people’s cognition beyond what might appear comfortable and necessary. Similarly to other forms of meta-information, it can indirectly reveal too much about processes and the datasets, information, and knowledge it is supposed to describe. Risks are apparent with, for example, various types of archival records, health information, and knowledge belonging or relating to vulnerable communities.

The studies of algorithmic accountability and legal ramifications of paradata point succinctly to difficulties to know whose interests paradata serve now and in the future. With automated paradata and using paradata for advancing algorithmic accountability, an obvious but admittedly, a convoluted question is what ethical questions are relevant when describing human involvement in processes and to what extent they are also relevant when describing automated or machine supported processes. Further, by increasing transparency of specific aspects of practices and processes, paradata might simultaneously hide and blackbox others either by accident or on purpose. Similarly to misinformation, there can also be non-truthful or even malign motivations to create paradata to promote particular narratives, distort the public image of a certain process or practice, or to "paradata-wash" a truly messy and poorly designed and executed endeavour. However, like many forms of misinformation, paradata-washing does not need to be malicious. Similarly to how Ruokolainen and Widén (2020) remind of misinformation, the dividing line between paradata and "mis-paradata" is not sharp. It is to an equal extent produced in a particular social situation.

However, in spite of all the thinkable caveats, if responsible, paradata have a capacity to do good. It has an apparent capability to realign some of the very fundamental "lines of accountability" (Guston, 1999) of how processes and practices become and remain dependable. By increasing transparency, both as form of codified knowledge and a lens, paradata can increase trust, shared understanding, equitability, and transparency. In attempts to decolonise data, paradata can tell about how they were collected, what wrongdoings were committed that could be repaired but also who was represented and to what degree a dataset actually is and is not biased. While this might entail an in-depth exploration of practices and processes, using paradata as a mere reminder to avoid overly simplistic conclusions can sometimes be enough. Similarly, even if the contemporary ideas of transparency and the aims of the Open movement are interpreted in terms of a close to unrestricted access to information, it is important to consider that there is nothing inherent in paradata suggesting that everything needs to be released or to be directly exploitable by others—whether they are multinationals, governments, individuals, or communities. Archival secrecy, privacy policies, patents, and the provisions in data sharing principles to make data as "open as possible and as closed as necessary" (Wilkinson et al., 2016) are all examples of how paradata can be generated but kept responsibly closed whenever needed.

Paradata ethics and its relation to general information and knowledge management ethics is without any doubt an important line of future research and practical attention. Incontestably, with ethics as well as with keeping up with paradata in general, the most important point of departure is to acknowledge the need of critical reflection. Like transparency is not given with or without paradata, the meritoriousness of paradata or any specific idea or form of transparency is equally little self-evident. Schenk and Reuß (chapter "Paradata in Surveys") point to the on-going debate on the requirement of informed consent about paradata collection and the limits of their confidentiality. Enqvist (chapter "Paradata as a Tool for Legal Analysis: Utilising Data-on-Data Related Processes") and Trace and Hodges (chapter "The Role of Paradata in Algorithmic Accountability") stress how extensive transparency easily conflicts with privacy. This is not least the case when legacy data is repurposed for new uses like in Cohen and colleagues' work of modelling of emergency service communications in chapter "Paradata for Digitization Processes and Digital Scholarly Editions". Often it is possible reach a trade-off but sometimes it is difficult to draw a line between what paradata are reasonable to keep and what needs to be discarded. In the current volume, survey studies and human subject research provide example of by definition benevolent activity that epitomises much of the intricacy of balancing between process transparency and keeping involved individuals non-identifiable. Comparable and even greater challenges are apparent in many other fields from healthcare to public and private security.

6 Future Perspectives on Paradata

It is obvious that with this volume we have only scratched the surface of the theory, practice, and implications of paradata for information and knowledge management and for the diverse settings where it is applied. The conceptual field is still very open, so is the field of practice. With perhaps the exception of survey research, the notion of paradata is still very much in the making without consolidated theory and practices of producing and exploiting it neither as a theoretical lens nor descriptive resource. However, the chapters throughout this volume show that it is stabilising, especially in fields like heritage visualisation, archivistics and artificial intelligence, systematising implications of the concept and its concrete instantiations. The forms and formats of paradata and the conceptual understanding of what the concept and paradata entail is a challenge that is approached in different disciplines from varying angles. For a legal scholar, paradata can work as a pedagogical concept, for survey researcher it is an established part of the scholarly toolbox, and for many fields of practice and scholarship it comes with a promise of shedding light and emphasis on matters that might have previously been under the radar. There is much work to be done to frame what counts as paradata and where to draw a line between paradata and other concepts. While the usefulness of theoretical uniformity can be debated, there is room for contrasting and comparing different theoretical understandings of paradata, finding synergies and complementarities.

The chapters shed light to different ways of collecting, extracting, creating, and curating paradata. Similarly to the need for a continuing theoretical discussion, there is a methodological discussion to be had. Relevant questions include, for instance, how interviews can be used to generate paradata in comparison with the on-going discussion about ethnography of fieldwork in archaeology. Another key line of future inquiry pertains to expectations and implications. Paradata is a concept that holds a great promise in helping to deliver the desired outcomes envisioned in the Open movement (European Commission, 2016) and the many of the contemporary societal aspirations for transparency, accountability and effective, responsible and equitable sharing and use of information and knowledge.

Finally, before departing to pursue future research, we feel that it is relevant to highlight two aspects of paradata that form a red thread through the texts in this volume. Creating paradata is an integral part of scientific method and in a broader sense, systematic knowledge-making. From the perspective of information and knowledge management, it is similarly an integral part of understanding and knowing what is managed. Another similarly crucial aspect of paradata is that it quite apparently changes the epistemological base for information and knowledge management. Paradata turns attention to what Prusak described as "thoroughly adopted" and invisible knowledge of processes (Prusak, 2001, p. 1006). Rather than being content with managing acontextual dichotomously true or false knowledge-things, paradata direct awareness to managing knowledge and information with history and future. With paradata we need and want to know how knowledge came into being.

References

Bunn, S., Bruun, M. H., Wahlberg, A., Douglas-Jones, R., Hasse, C., Hoeyer, K., Kristensen, D. B., & Winthereik, B. R. (Eds.) (2022). Technology as skill in handwork and craft: Basketwork and handweaving. In *Palgrave handbook of the anthropology of technology* (pp. 61–83). Palgrave Macmillan.

Cameron, S., Franks, P., & Hamidzadeh, B. (2023). Positioning paradata: A conceptual frame for AI processual documentation in archives and recordkeeping contexts. *Journal of Computing and Cultural Heritage, 3594728*. https://doi.org/10.1145/3594728

Denard, H. (2012). A new introduction to the London Charter. In H. Denard & D. Baker (Eds.), *Bentkowska-Kafel Anna* (pp. 57–71). Paradata and transparency in virtual heritage. Ashgate.

Edwards, R., Goodwin, J., O'Connor, H., & Phoenix, A. (2017). *Working with paradata, marginalia and fieldnotes*. Edward Elgar.

European Commission. (2016). *Open innovation, open science, open to the world | Shaping Europe's digital future.*

Guston, D. H. (1999). Stabilizing the boundary between US politics and science: The rôle of the office of technology transfer as a boundary organization. *Social Studies of Science, 29*(1), 87–111. https://doi.org/10.1177/030631299029001004

Handzic, M. (2004). *Knowledge management: through the technology glass*. World Scientific.

Hess, D. J. (2005). Technology- and product-oriented movements: Approximating social movement studies and science and technology studies. *Science, Technology, & Human Values, 30*(4), 515–535. https://doi.org/10.1177/0162243905276499

Huggett, J. (2020). Capturing the silences in digital archaeological knowledge. *Information-an International Interdisciplinary Journal, 11*(5), 278. https://doi.org/10.3390/info11050278

Huvila, I. (2022). Improving the usefulness of research data with better paradata. *Open Information Science, 6*, 28–48. https://doi.org/10.1515/opis-2022-0129

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.

Polanyi, M. (2009). *The tacit dimension (Revised)*. The University of Chicago Press.

Prusak, L. (2001). Where did knowledge management come from? *IBM Systems Journal, 40*(4), 1002–1007. https://doi.org/10.1147/sj.404.01002

Ruokolainen, H., & Widén, G. (2020). Conceptualising misinformation in the context of asylum seekers. *Information Processing and Management, 57*(3). https://doi.org/10.1016/j.ipm.2019.102127

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18

Isto Huvila PhD, is professor in information studies at the Department of ALM at Uppsala University in Sweden. He received an MA degree in cultural history at the University of Turku in 2002 and a PhD degree in information studies at Åbo Akademi University (Turku, Finland) in 2006. Huvila was chairing the recently closed COST Action ARKWORK and is directing the ERC funded research project CAPTURE. His primary areas of research include information and knowledge management, information work, knowledge organisation, documentation, and social and participatory information practices.

Lisa Andersson Ph.D., works as a researcher at the Department of ALM at Uppsala University in Sweden. She received her MA degree in library and information science in 2011, and her doctoral degree in library and information science in 2017, both at the Department of ALM at Uppsala University in Sweden. Her research focuses on data and information management including research data and information management systems, knowledge organisation and data descriptions, data publishing and use. Andersson has published in library and information science journals but also in cross-disciplinary journals from the fields of archaeology and digital humanities.

Olle Sköld Ph.D., works as a Senior Lecturer at the Department of ALM at Uppsala University in Sweden. He received his MA degree in archival studies in 2010 at Lund University, and a doctoral degree in information studies in 2018 at Uppsala University. His research is characterised by a broad interest in the ALM field, knowledge organisation and production, research data creation and use, and digital humanities. Sköld has published in information studies journals including the Journal of the Association for Information Science and Technology, Journal of Documentation and Information Research.