DE GRUYTER

Nicola Reggiani (Ed.)

DIGITAL PAPYROLOGY III

THE DIGITAL CRITICAL EDITION OF GREEK PAPYRI: ISSUES, PROJECTS, AND PERSPECTIVES

Digital Papyrology III

Digital Papyrology III

The Digital Critical Edition of Greek Papyri:
Issues, Projects, and Perspectives

Edited by Nicola Reggiani

DE GRUYTER

The present volume is funded by the Department of Humanities, Social Studies, and Cultural Enterprises (DUSIC) of the University of Parma in the framework of the PRIN National Project “Greek and Latin Literary Papyri from Graeco-Roman Egypt (4th BC – 7th AD): Texts, Contexts, Readers” (LitPapArs, Principal Investigator Professor Lucio Del Corso, University of Salerno), local research unit (coordinator Professor Nicola Reggiani, University of Parma).

ISBN 978-3-11-107013-1
e-ISBN (PDF) 978-3-11-107016-2
e-ISBN (EPUB) 978-3-11-107055-1
DOI https://doi.org/10.1515/9783111070162

Library of Congress Control Number: 2017299846

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Preface: The Paradox of Papyrology and the Digital Paradigm

Papyrology is a discipline of many paradoxes and contradictions. Paradoxical is the very core of the evidence it relies on: hundreds of thousands of smaller or larger 'textual items', now scattered across dozens of collections around the world despite being found together, and mostly coming from small, peripheral towns. As a consequence, papyri do not help us so much in reconstructing the long-dreamed Library of Alexandria and its lost books, but they show us the reading choices of a clerk in Karanis, or the daily routines of a school teacher in the even farther Trimithis; and, while major historical events are rarely mentioned in them, the plethora of surviving documents offers us vivid fragments of the daily lives of ordinary people, who are normally destined to disappear.

The efforts to decipher and understand such tantalizing texts, and to use them to turn sporadic pieces of information into a coherent historical picture, had a peculiar side effect: the paradoxical nature of the extant evidence urged papyrologists to explore new quantitative approaches, in order to sort and serialize the data, and this soon led to the development of databases, libraries of digital images, and further IT tools in general. Papyrology has become intrinsically 'digital', and Nicola Reggiani has already reconstructed the history of this evolution in a seminal volume, not only providing a survey of how IT technology has improved the possibility to decipher and understand papyrological material, but also exploring the methodological and epistemic consequences of this new paradigm.[1] Even a few years later, the pace of evolution seems so fast to exceed the best expectations. The *Vesuvius Challenge*, relying on a ground-breaking image acquisition technique (X-ray tomography), AI software, and a very high budget, has demonstrated how collaborative efforts of computer scientists and papyrologists can achieve previously unimaginable results: reading a carbonized scroll without unfolding (and breaking) it, and producing, in a few months, a masterly *editio princeps* of 16, previously sealed, half-columns from a lost Epicurean treatise.[2]

The new Herculaneum discoveries can be placed within a broader frame, which is clear from the essays collected in this volume. Besides the textual discoveries that IT technologies will allow, digital papyrology has now entered a new phase: from documenting extant evidence and arranging it into a coherent order, to offering a new way

1 See esp. N. Reggiani, *Digital Papyrology I. Methods, Tools, and Trends*, Berlin – Boston 2017, later updated as N. Reggiani, *La papirologia digitale. Prospettiva storico-critica e sviluppi metodologici*, Parma 2019.

2 F. Nicolardi – D. Delattre – G. Del Mastro – R. Fowler – R. Janko, *The Final Columns of PHerc.Paris. 4 Revealed through Virtual Unwrapping*, Cronache Ercolanesi 54 (2024), 9–27. The achievements of the *Vesuvius Challenge* had a vast echo on newspapers and media all around the world; full information is provided by the project website (https://scrollprize.org). The first steps of the application of X-ray tomography to Herculaneum papyri are described also in Reggiani, *Digital Papyrology I*, 148–9.

https://doi.org/10.1515/9783111070162-001

to analyze and interpret it from a *qualitative* point of view, so to get a better, more complex and refined understanding of the texts and their original contexts.

The three sections of this volume reflect the main areas in which the new hermeneutic efforts are directed: the characteristics of digital critical editions, between the inherent 'polysemy' of XML and other encoding languages and the need to offer a stable, authoritative text – the ultimate goal of any philological effort –, especially when the manuscript source can be reconstructed only starting from electronic imaging, as in the case of the latest Herculaneum discoveries; the possible interactions between the corpora of encoded papyri and computational linguistics; the new frontiers of 'digital palaeography' and the different approaches that new projects are following to answer debated questions, such as the possibility of dating scripts and handwritings on a firm foundation, and connecting them to specific individuals. Reading the pages of this book – and recalling the discussions that arose when the papers were first presented – one has the impression that it is not utopian to hope for future advancements and exciting discoveries, through a mix of new technologies and more traditional 'philological' knowledge. But they will be achieved only if the study of multifaceted phenomena is not reduced to mere quantitative data: the accumulation of information does not guarantee *per se* a deeper knowledge. The challenge of the 'digital approach' to papyrology, and more generally to the study of the ancient world, will be not only reading what is now illegible, but also providing a new, comprehensive way to represent complex models, and finding connections that would otherwise be lost, within a truly historical perspective.

It is a pleasure, indeed, that the reflections contained in this book originated in the framework of the PRIN 2017 Project "Greek and Latin Literary Papyri from Graeco-Roman and Late Antique Arsinoites", which aimed to start from a specific case-study in order to explore wider, socio-cultural perspectives, developing both 'traditional' books and new digital resources, as the database *LitPapArs*, which will be available online during 2025. We can just hope that the wealth of knowledge and information offered by the texts from Graeco-Roman Egypt will become more and more available also outside the small number of militant papyrologists: and if this is happening, much of the credit is due to digital papyrology.

Lucio Del Corso

Foreword

The present volume is the natural continuation of the two preceding works on Digital Papyrology,[1] not only in the mere numbering sequence but also in the epistemological and methodological development.

Since its beginnings, Digital Papyrology has been changing the way of considering its object of study as well as the very object of study itself, which has become a virtual *avatar* of the physical papyrus – and not only a bidimensional editorial representation of it –, a meta-papyrus activating a network of cognitive interconnections between data and metadata. From this perspective, the ancient papyrus text can be perceived as a hypertext – not a fixed, static, stable text but an interactive and open system involving the ancient author, the ancient reader, the modern scholar.

Today, the digital environments allow for envisioning a critical edition that is not a simple reproduction of a scholarly idea of text, but a careful representation of the original system of cognitive interactions. Indeed, the title of the volume also winks at the so-called Web 3.0, usually intended as integrated semantic web, which deploys database-like structures rather than hypertextual pages. This evolution corresponds also to the evolution of the papyrological databanks, from simple repositories of texts and metadata to editorial representations and – eventually – integrated editorial platforms.

The current need to go beyond the digital encoding of the pure and simple papyrus texts on Papyri.info (which is, however, a semantic markup, though focused on the main papyrological and editorial information about the text itself) is well represented by pathbreaking projects currently in progress, all dealing with the digital annotation of particular features of the papyri. This volume is precisely devoted to such projects, which have developed in the years following the publication of *Digital Papyrology II* (2018).

The chapters stem from the papers presented by a worldwide group of project-leading scholars at the international conference "Digital Papyrology 3.0 – Digital Encoding and Critical Edition of Greek Papyri: perspectives and progress", held at the University of Parma on May 30–31, 2022.[2] The articulation of the volume follows the same logical arrangement of the conference sessions. The first part is devoted to general theoretical

1 N. Reggiani, *Digital Papyrology I. Methods, Tools, and Trends*, Berlin – Boston 2017 (updated in N. Reggiani, *La papirologia digitale. Prospettiva storico-critica e sviluppi metodologici*, Parma 2019); N. Reggiani (ed.), *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, Berlin – Boston 2018.

2 http://www.papirologia.unipr.it/eventi/dp3 [last access 31.8.24]. Both that conference and this volume fall within the framework of the PRIN 2017 project "Greek and Latin Literary Papyri from Graeco-Roman and Late Antique Fayum: Texts, Contexts, Readers" (Principal Investigator: Lucio Del Corso, University of Salerno), local research unit at the University of Parma (coordinator: Nicola Reggiani). Exactly due to the connection with this project, I am most honoured to host the introductory words of Lucio Del Corso, whom I heartfeltly thank.

https://doi.org/10.1515/9783111070162-002

questions on the digital critical edition and its virtual environment, not necessarily from the viewpoint of Papyrology only, but from the more comprehensive standpoint of the Digital Humanities. Indeed, since we are increasingly virtualizing our objects of study and our methodologies, it is important to reflect on the media that are going to host and shape our future work. The second section deals with projects involving the core data of the papyri: the texts, their digital encoding, and the linguistic applications to the text studies. The third part is focused on projects dealing with the surrounding characteristics of papyrus texts: material features, visual semiotics, structural textures, palaeography. The last chapters – involving automatic recognition of scribal handwritings – build a bridge towards the next step in Digital Papyrology: the applications of Artificial Intelligence and Machine Learning systems, which are not the main focus of this volume, but are necessarily faced by the most recent electronic developments of the discipline.[3]

Human necessities brought the conference papers to their written format after two years,[4] but the purposes and the significance of the project have not changed. The hope is for a future possible interconnection of all these annotation levels towards a true digital critical edition of papyrus texts.

Nicola Reggiani

Parma, June 10, 2024

3 See N. Reggiani, *The Artificial Papyrologist at Work*, in *Decoding Cultural Heritage: A Critical Dissection and Taxonomy of Human Creativity through Digital Tools*, ed. by F. Moral-Andrés, E. Merino-Gomez, and P. Reviriego, Cham, 123–36.

4 I would like to express my gratitude for the professional help and extreme patience demonstrated by the editors at De Gruyter, who followed the various stages of composition of this volume – namely, Mirko Vonderstein, Florian Ruppenstein, Anne Hiller, Jessica Bartz, Torben Behm.

Contents

Part I: **Theoretical Frameworks of the Digital Critical Editions**

Nicola Reggiani

The Digital Critical Edition of the Papyri: Topics, Issues and Perspectives

1 The digital textual criticism of the papyri

> Whether the papyrologist is reading a papyrus for the first time or reviewing an earlier transcription, every step that he takes, from the first recognition of letters, words, or phrases to his final explanation of the completed text, requires the exercise of critical power.
>
> Herbert C. Youtie[1]

Ancient textual transmission has been traditionally regarded with a philological approach. The concept of critical edition frames the philological aspiration to the *constitutio textus*, the reconstruction and re-establishment of the original form (the source, 'archetype' or *Urtext*) of a text as most exactly as possible, i.e., the most possible corresponding to the author's concept, on the ground of the collation of the different extant direct and indirect testimonies of that text (the 'witnesses'), which may often present variants that need to be critically compared and sifted in order to fetch the original source.[2] As a consequence, the typical printed layout of a traditional critical edition consists of a section devoted to the ideally correct text and another section reserved to the *apparatus criticus*, which registers all the possible attested variants and any editorial annotations. This concept, therefore, is clearly based on a text abstraction, i.e., a hopefully stable representation of a scholar's more or less reliable opinion on the text, 'artificially' reconstructed from actual, frequently discordant items.

The whole chapter falls into the framework of the PRIN 2017 National Project "Greek and Latin Literary Papyri from Graeco-Roman and Late Antique Fayum (4th BCE–7th CE): Texts, Contexts, Readers" (Principal Investigator: Lucio Del Corso, University of Salerno), Local Research Unit at the University of Parma (coordinator: Nicola Reggiani). This section develops two hitherto unpublished presentations: "Towards a Digital Criticism of Ancient Greek Papyri", delivered at the 11th Celtic Classics Conference, University of St Andrews, on July 13, 2018, and "Towards a Digital Criticism of Greek Papyrus Texts", delivered at the Universität Würzburg on January 31, 2019, within the framework of the DAAD-MIUR project "Ekdosis: Digitizing Literary and Paraliterary Papyri" (Universities of Parma and Würzburg, Principal Investigators: Massimo Magnani and Holger Essler). All hyperlinks last accessed on 31.8.24.

1 Youtie 1974, 6.

2 See Maas 1960; Reynolds – Wilson 1991, 207–41; Timpanaro 2004; Braccini 2017.

https://doi.org/10.1515/9783111070162-003

When moving to the electronic environments, most of the critical editions look deeply indebted to such a traditional framework. We can recognize three different stages stemming from the very same philological root: (1) a digital reproduction of a printed edition (e.g., the scan of a printed edition); (2) a digital transcription of a printed edition (e.g., a Word or a PDF file typed according to the traditional editorial conventions); (3) a digital transcoding of a printed edition. This is, for instance, the strategy followed by Papyri.info, the ultimate papyrological databank, which stores papyrus critical editions transcribed in Unicode characters and transcoded into two parallel markup languages – XML, Leiden+ –, yet still similar to a printed traditional critical edition in concept and display:[3] a main text and an apparatus collecting editorial interventions[4] – even in the case of born-digital editions.[5]

The limits of a pure stemmatological philological methodology have been remarked since several years, with reference to more fluid forms of transmissions involving texts that evolve and transform to suit changing needs and circumstances, so that the idea of a unique and fixed 'original' is felt as rather uncomfortable.[6] As a consequence, while an editor's *auctoritas* supports the philological choices that lead to obtain an 'authorial' text which is reputed to be the most correct and genuine, we must not forget three important deviations from the 'ideal world': (1) the possible existence of several original or 'authorial' versions; (2) the living nature of technical texts;[7] (3) the viewpoint of the "copyist as an author,"[8] which pinpoints the fact that what actually found circulation, diffusion, and reception in antiquity were the single, material copies of the text, influenced in various ways by the scribes who penned them.

Therefore, enterprises like the *Homer Multitext Project* (HMT) started envisaging a different digital editorial approach, involving a text which is in fact a multitext, a network of multiple editions interconnected to each other, rather than a traditional fixed structure of text and apparatus criticus.[9] The focus of such projects is not on the original text but on its very transmission, giving emphasis to the single instances of the transmission itself. It is clear that this is a risky operation, which could lead to excessive devaluations, in the name of a possible "agnosticism,"[10] in that the single testimonies are simply juxtaposed in parallel – as the HMT's logo proudly announces: "As many Homers as you please." Nonetheless, the HMT holds the merit of stressing the value of the single identity of each textual

3 See http://digitalpapyrology.blogspot.it/2011/03/new-in-ddbdp.html: "we have taken the first major steps toward bringing the DDbDP's apparatus criticus conventions more closely into line with current practice."

4 See Reggiani 2017, 222–40, and 2019a, 292–316.

5 On born-digital papyrus editions see Berkes 2018.

6 See Pasquali 1988; Reynolds – Wilson 1991, 234–7 and 288–92; Braccini 2017, 115–21.

7 See Reggiani 2018a and below, §3.

8 Canfora 2002.

9 See Nagy 2010; Magnani 2018, 94–9.

10 Bodard – Garcés 2009, 96 n. 31.

incarnation, being the direct digital heir of Albert Lord's famous opinion about Homeric transmission, which can be resumed in the keyword 'multiformity':

> Our real difficulty arises from the fact that, unlike the oral poet, we are not accustomed to thinking in terms of fluidity. We find it difficult to grasp something that is multiform. It seems to us necessary to construct an ideal text or to seek an original, and we remain dissatisfied with an ever-changing phenomenon. I believe that once we know the facts of oral composition we must cease trying to find an original of any traditional song. From one point of view each performance is an original.[11]

Papyrology allows for a privileged perspective on the issue at stake. Indeed, it has always been coping with an adventurous textual situation, dealing with fragmentary texts and idiosyncratic utterances (scribal personal uses, linguistic change and variation), and being particularly interested in the scribal and material phenomenology of textual development and transmission. Traditionally, Papyrology is a philological discipline,[12] focused on texts and their critical reconstruction, but the fact that its objects of study – the texts preserved on papyrus and other everyday portable supports from Hellenistic, Roman, and Late Antique Egypt (and not only) – are direct and almost unique witnesses, original and direct expressions of the texts as they were produced, circulated, and utilized, challenges its own philological nature and brings to the front the issue of the concreteness of the texts and of their transmission.

Not by chance, in recent times we are facing a profoundly renewed scholarly interest toward the text as a material product,[13] or better toward the indissoluble relations between the text as a cognitive product of scribal activity and the actual materiality of the writing support, which concur to shape what can be described as the "phenomenology of the text."[14] From this viewpoint, the text still assumes an undoubtedly central position, but it is not an 'abstract' or 'absolute' text, it is the text as it was thought and produced by the ancient scribe or copyist, and as it was actually enjoyed in the original context of circulation and transmission – a text immersed in a network of cognitive strategies.[15] Such observations are certainly valid for the documentary texts, which are unique products of the scribal activity, without archetypes and stemmas, not even in the cases of duplicate documents or authorial revision stages, for which we may at least resort to the so-called 'genetic criticism'.[16] Yet they remain valid also for the literary texts, though provided of its own preceding and subsequent tradition, in which the papyrus however is a unique expression of an individual and characteristic writing act:

11 Lord 1960, 100.

12 See Hanson 2002, 193; Schubert 2009, 197.

13 See e.g. Meier – Ott – Sauer 2015; Hoogendijk – van Gompel 2018; Sarri 2018; Ast – Choat – Cromwell *et al.* 2021; Reggiani 2024a. See also the chapters authored by Daniel Riaño and Serena Causo in this volume

14 Reggiani 2018a, 7.

15 Examples with further bibliography are offered by the projects presented in this volume. See the Foreword for a general overview.

16 On genetic criticism applied to papyri see Cribiore 2019.

> Il pregio della testimonianza papiracea è di conservare l'aderenza formale e contenutistica dello scritto alla sua destinazione e di suggerire la dinamica del rapporto di composizione e copia con quello di fruizione del contenuto, quale che fosse la dignità del genere cui è appartenuto.[17]

The preceding observations bring us to consider the main limits of traditional printed papyrus editions. First of all, papyrus editions feature one scholar's fixed opinion about a text. Due to the strong materiality of the items – texts are fragmented and broken, and exhibit very peculiar handwritings and writing strategies – their editions are subject to corrections and updates, often difficult and/or slow to handle in a traditional paper environment. Papyrology is admittedly a "discipline in flux"[18] and needs to undergo a 'liquid' philology that envisages editorial changes and scholarly progress through time.[19] This issue is now addressed by the history log of Papyri.info, which records any editorial emendation to the encoded texts.[20]

Second, traditional papyrus editions aim at establishing an original / correct text. Nevertheless, a papyrus *Urtext* does in fact not exist. Each piece of text is a unique and direct testimony of a scribal utterance and deserves a careful attention to the scribal phenomenology.[21] As eventually stated by Herbert Youtie, who attempted a theorizing definition of textual criticism applied to the papyri, a papyrus critical edition must be a reliable representation of a papyrus text: the papyrus is "the final arbiter"[22] of any interpretation. Editorial interpretation is still a fundamental step (e.g., in supplying the text lost in the material lacunas, or in correctly recognising and dividing the words written in *scriptio continua*) but the editor must act not more than "a skilled reader" and "a learned copyist" aiming at the "coherent meaning" of the edited text.[23]

The editorial representations of the texts can vary from the diplomatic transcription of what remains on the writing surface (which was one of the main concerns of the early databases, later abandoned) to a 'hybrid' edition that tries to preserve the restoration of a text as close as possible to a 'regular' original alongside the recording of variant readings: for example, the early Duke Databank of Documentary Papyri, which encoded the 'normalized' / 'regularized' / 'correct(ed)' words in the main text and the 'variant' forms – as written on the original papyrus – adjacent to the former, marked

17 Andorlini 1993, 462. ["The value of the papyrus witnesses lies in preserving the adherence – as regards form and content – of the written text to its intended purpose and in suggesting the dynamics of the relationship between the composition and copying process and the use of the content, regardless of the status of the genre to which it belonged."]

18 Hanson 2002; see also Youtie 1963, 31–2, and 1974, 6–7; Schubert 2009, 212–3.

19 See Reggiani 2017, 264, and 2019a, 349. The concept of 'liquidity' has been theorized by Zygmunt Bauman, emphasizing the fact of change in the contemporary society (see e.g. Bauman 2000; 2007; 2011).

20 See Reggiani 2017, 233 and 265; 2019a, 306 and 350.

21 See Youtie 1974, 13–15.

22 Youtie 1974, 25.

23 Youtie 1974, 2, 22, and 16, respectively.

with special notation (Fig. 2),[24] which later became an apparatus note followed by 'Pap.' (Fig. 3). The solution adopted today by Papyri.info – original reading in the text, normalization/correction in the apparatus (Fig. 4) – is appropriate in regards of the rendering of the original phenomenology of the text, but is still indebted to an editorial criticism that considers the 'variant' as a deviation from a standard 'archetype', to be normalized not only visually – by displaying an imperative *l(ege)* before the 'normalization' in apparatus – but also semantically, using the XML tag `<reg>` for 'regularization'.[25] While this may be suitable for scribal errors proper (but what is an error?), it poses some discomfort regarding orthographic/linguistic variants, which are increasingly considered important cultural factors rather than deviations from a theoretical norm that, in fact, did not exist.[26] The discourse is not purely theoretical: in a digital environment, it affects search functions, since – for example –, to date, the Papyrological Navigator of Papyri.info cannot perform proximity searches involving words in the apparatus.

When we turn to literary papyri, the problem is even more complex because – besides linguistic variants and mechanical mistakes – we frequently find philological variants, with the need to establish whether the reading of the papyrus is in total or partial agreement with the manuscripts or represents a completely new variant. We can also find scribal variants, where the scribe himself notes two different versions of the same word.[27] In a traditional critical apparatus format, we must decide which text is to be considered 'normal' or 'regular' and which constitutes a secondary reading. But what if the scribe judged a reading different from what we derive from the philology of the manuscript tradition to be correct? And what if the papyri attest to a minority variant that is nevertheless undoubtedly ancient? 'Paraliterary' texts present even more difficult situations: in technical writings, the textual transmission follows the tortuous paths of oral teaching, practical learning, and enrichment from the individual experience of each specialist. Thus, we find that a supposed 'archetype' – e.g., a medical prescription – often evolves into different versions: quotations, comments, summaries, revisions, personal reinterpretations, and contingent variants connected to practical use.[28]

This is an extremely fluid situation, and while printed paper can hardly respond to complex editorial claims for both practical and theoretical reasons, a digital environment is particularly suitable to that double purpose: the digital critical edition must be a workspace where the text is not a fixed boundary but a fluid set of information open to corrections and updates, and where it can be represented at best in its material phenomenology and in the complex network of cognitive strategies that produced and utilized it.

24 See Reggiani 2017, 214–7, and 2019a, 282–5.

25 In general, on the issue of encoding linguistic variation in the papyri see Stolk 2018. See also Reggiani 2018a and 2019b.

26 See below, §4.

27 See Reggiani 2019b and 2019c.

28 See Reggiani 2019d and below, §§3–4.

This does not mean an acritical or agnostic multitextual juxtaposition, but a critical representation of an actual stage of text transmission. Of course, central in the digital critical workflow is still the role of papyrologists: despite the rapid development of Artificial Intelligence systems, reading and editing a papyrus is still a matter of human criticism in terms of mental / intellectual process.[29] The papyrologist still holds the responsibility of being an "artificer of facts,"[30] which means that the critical edition (s)he produces becomes the true source of further studies and investigations. Since we cope with ever-changing facts, the aim of digital criticism is to keep traces of all of them in an open and 'liquid' edition that is a faithful representation of the papyrus text and all its possible cognitive networks, and that configures itself as a further step in the textual transmission rather than a fixed *Urtext*.[31]

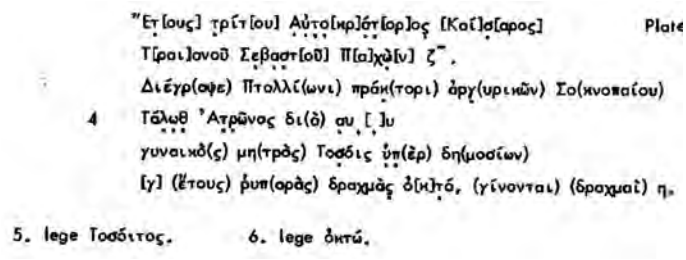

Ἔτ[ους] τρίτ[ου] Αὐτο[κρ]άτ[ορ]ος [Καί]σ[αρος] Plate
Τ[ραι]ανοῦ Σεβαστ[οῦ] Π[α]χὼ[ν] ζ.
Διέγρ(αφε) Πτολλί(ωνι) πράκ(τορι) ἀργ(υρικῶν) Σο(κνοπαίου)
4 Τάλωθ Ἀτρῶνος δι(ὰ) αυ [ ]υ
γυναικὸ(ς) μη(τρὸς) Τοσδις ὑπ(ὲρ) δη(μοσίων)
[γ] (ἔτους) ρυπ(αρὰς) δραχμὰς ὀ[κ]τό, (γίνονται) (δραχμαὶ) η.
5. lege Τοσόιτος. 6. lege ὀκτώ.

Fig. 1: Printed edition of P.Coll.Youtie I 33 (see below, §2).

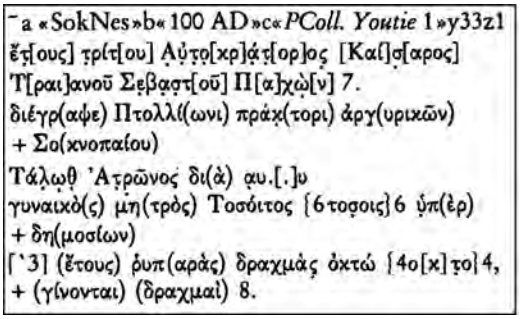
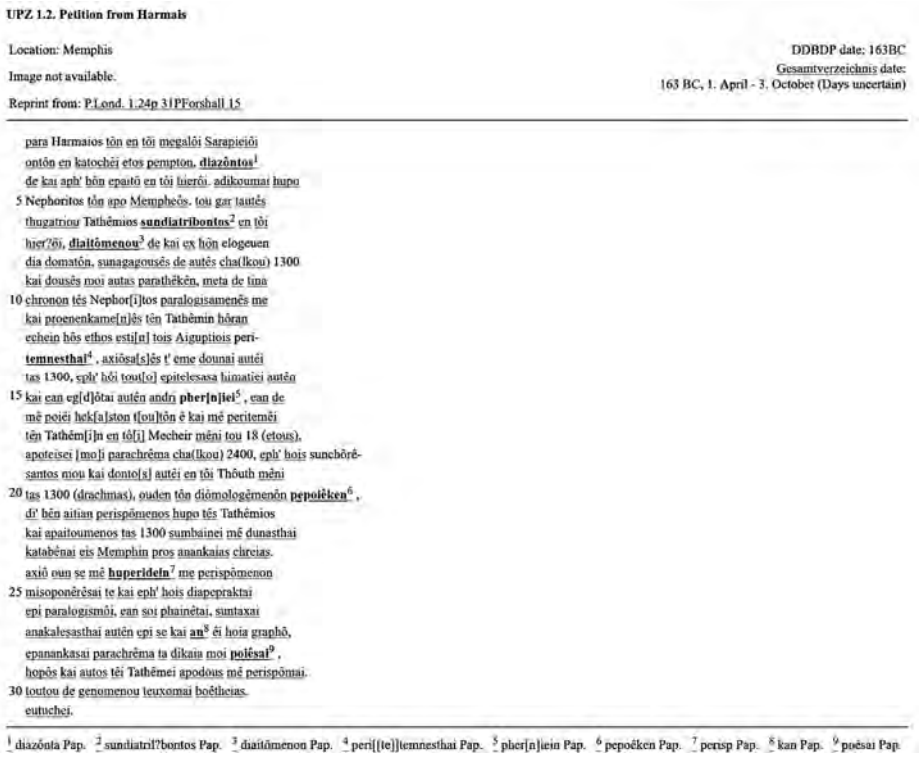

¯a «SokNes»b«100 AD»c«*PColl. Youtie* 1»y33z1
ἔτ[ους] τρίτ[ου] Αὐτ[κρ]άτ[ορ]ος [Καί]σ[αρος]
Τ[ραι]ανοῦ Σεβαστ[οῦ] Π[α]χὼ[ν] 7.
διέγρ(αφε) Πτολλί(ωνι) πράκ(τορι) ἀργ(υρικῶν)
+ Σο(κνοπαίου)
Τάλωθ Ἀτρῶνος δι(ὰ) αυ.[.]υ
γυναικὸ(ς) μη(τρὸς) Τοσόιτος {6τοσδις}6 ὑπ(ὲρ)
+ δη(μοσίων)
⌈`3] (ἔτους) ρυπ(αρὰς) δραχμὰς ὀκτώ {4ο[κ]το}4,
+ (γίνονται) (δραχμαὶ) 8.

Fig. 2: P.Coll.Youtie I 33 (see Fig. 1) in the early Duke Databank of Documentary Papyri (Willis 1984, 170–1).

29 See Reggiani 2024b.

30 Youtie 1963; see Youtie 1966, 257–8, and 1974, 23.

31 See Reggiani 2022a and below, §5.



UPZ 1.2. Petition from Harmais

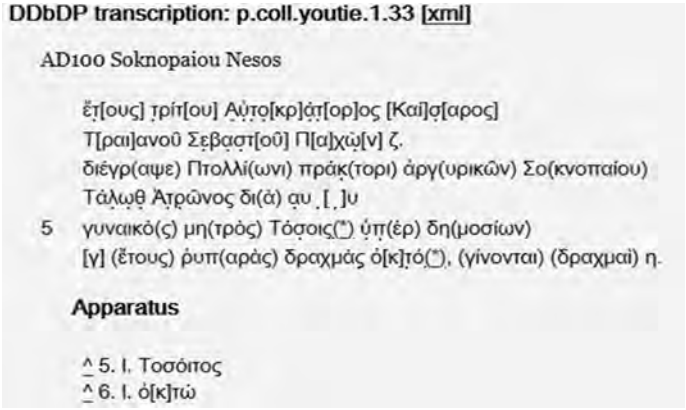
Location: Memphis
Image not available.
Reprint from: P.Lond. 1.24p 31PForshall 15

DDBDP date: 163BC
Gesamtverzeichnis date:
163 BC, 1. April - 3. October (Days uncertain)

para Harmaios tôn en tôi megalôi Sarapieiôi
ontôn en katochêi etos pemptôn, **diazôntos**[1]
de kai aph' hôn epastô en tôi hierôi. adikoumai hupo
Nephoritos tôn apo Mempheôs. tou gar tautês
thugatriou Tathêmios **sundiatribontos**[2] en tôi
hier?ôi, **diaitômenou**[3] de kai ex hôn elogeuen
dia domatôn, sunagagousês de autês cha(lkou) 1300
kai dousês moi autas parathêkên, meta de tina
chronon tês Nephor[i]tos paralogisamenês me
kai proenenkame[n]ês tên Tathêmin hôran
echein hôs ethos esti[n] tois Aiguptiois peri-
temnesthai[4], axiôsa[s]ês t' eme dounai autêi
tas 1300, eph' hôi tou[to] epitelesasa himatiei autên
kai ean eg[d]ôtai autên andri **pher[n]iei**[5], ean de
mê poiêi hek[a]ston t[ou]tôn ê kai mê peritemêi
tên Tathêm[i]n en tôi[i] Mecheir mêni tou 18 (etous),
apoteisei [mo]i parachrêma cha(lkou) 2400, eph' hois sunchôrê-
santos mou kai donto[s] autêi en tôi Thôuth mêni
tas 1300 (drachmas), ouden tôn diomologêmenôn **pepoiêken**[6],
di' hên aitian perispômenos hupo tês Tathêmios
kai apaitoumenos tas 1300 sumbainei mê dunasthai
katabênai eis Memphin pros anankaias chreias.
axiô oun se mê **huperidein**[7] me perispômenon
misoponêrêsai te kai eph' hois diapepraktai
epi paralogismôi, ean soi phainêtai, suntaxai
anakalesasthai autên epi se kai **an**[8] êi hoia graphô,
epanankasai parachrêma ta dikaia moi **poiêsai**[9],
hopôs kai autos têi Tathêmei apodous mê perispônai.
toutou de genomenou teuxomai boêtheias.
eutuchei.

[1] diazônta Pap. [2] sundiatrit?bontos Pap. [3] diaitômenon Pap. [4] peri[[te]]temnesthai Pap. [5] pher[n]iein Pap. [6] pepoêken Pap. [7] perisp Pap. [8] kan Pap. [9] poêsai Pap.

Fig. 3: UPZ I 2 (P.Lond. I 24 [TM 3393], petition, Memphis, 163 BC) in the online version of the Duke Databank of Documentary Papyri formerly hosted by the Perseus Digital Library (https://web.archive.org/web/20200006180602 24/http:/www.perseus.tufts.edu/cgi-bin/ptext?doc=Perseus:text:1999.05.0245). Unfortunately, the Wayback Machine service of Archive.org did not preserve a copy of the databank record of P.Coll.Youtie I 33.



DDbDP transcription: p.coll.youtie.1.33 [xml]

AD100 Soknopaiou Nesos

ἔτ[ους] τρίτ[ου] Αὐτο[κρ]ά̣τ[ορ]ος [Καί]σ̣[αρος]
Τ[ραι]ανοῦ Σεβα̣σ̣τ[οῦ] Π[α]χ̣ὼ̣[ν] ζ.
διέγρ(αψε) Πτολλί(ωνι) πράκ̣(τορι) ἀργ(υρικῶν) Σο(κνοπαίου)
Τάλ̣ω̣θ̣ Ἀτρ̣ῶνος δι(ὰ) α̣υ̣ [ ]υ
γυναικό(ς) μη(τρὸς) Τόσ̣οις(*) ὑ̣π̣(ὲρ) δη(μοσίων)
[γ] (ἔτους) ῥυπ(αρὰς) δραχμὰς ὁ[κ]τ̣ό̣(*), (γίνονται) (δραχμαὶ) η.

Apparatus

^ 5. l. Τοσόιτος
^ 6. l. ὁ[κ]τώ

Fig. 4: P.Coll.Youtie I 33 (see Fig. 1) in Papyri.info today (https://papyri.info/ddbdp/p.coll.youtie;1;33).

2 The materiality of the papyri in the digital age

> Non è la materia che genera il pensiero, è il pensiero che genera la materia.
>
> Giordano Bruno[32]

As introduced in the previous section, the papyrus text is not an abstract and ideal entity but an actual product of a writing act, where the text itself as carrier of linguistic meaning is necessarily intertwined with both the material object supporting it and the writing strategies deployed to represent its meaning.[33] The materiality I will consider in this section is therefore twofold: (1) the material support of the text; (2) the writing strategies of the text.

As regards the first aspect, Digital Papyrology has certainly increased and enhanced the perception of the materiality of the papyri. The first papyrus editions were seldom flanked by photographic reproduction, usually limited to the most important texts. The development of digital imaging has allowed to increase the possibility to access and examine the papyrus, often in an augmented way: plain digital pictures can be scaled, enhanced and manipulated, and further processed in three-dimensional models that better represent the materiality of the objects; complex reconstructions allow for virtual restorations of damaged or dispersed pieces, and even the virtual unwrapping of rolls difficult to handle; photographic shots in the non-visible wavelengths (infrared, ultraviolet, X-rays) can reveal invisible or ill-preserved ink traces.[34]

All of such strategies orbit around digital representations of the objects of papyrological studies, i.e. papyri and related materials. As was pointed out some time ago,[35] such representations – surrogates of the originals – may bear further uncertainties, beside those intrinsically embedded in fragmentary, damaged, abraded texts, mainly due to technical distortions coming from the adaptation of a material artefact to a digital medium. This poses a big caveat, especially related to the technical standards to be developed in order to ensure a proper digital representation of the objects. However, this is by no means aimed at diminishing the reliability of digital pictures, but just to make it sure that we all are aware that we are not dealing with the original, material objects, but with virtual artefacts, 'avatars' of the original pieces,[36] and that all the resources and the potentialities developed so far rely on this very fact.

32 ["It is not matter that generates thought, it is thought that generates matter."]

33 This section develops a hitherto unpublished talk, titled "The Materiality of the Greek Medical Papyri in the Digital Age", delivered at the Institut für Papyrologie, Universität Heidelberg, on April 12, 2018, in the frame of a Mercator Fellowship granted by the SFB Materiale Textkulturen, for which I am most grateful to Professor Andrea Jördens.

34 See Reggiani 2017, 137–61, and 2019a, 201–35; Fleischer 2021; Reggiani 2021.

35 Terras 2011.

36 The concept of 'avatar' is taken from Tarte 2016.

This leads us to a further step: the interpretive act embedded in the digitizing process, as investigated by Ségolène Tarte some years ago.[37] Several techniques and methodologies applicable to the digital representation of a papyrological object tend to reproduce the papyrologist's interpretive acts: Tarte brings the example of the shadow-stereo imaging of the stylus tablets, reproducing the actual different angles from which the researcher looks at the objects; but also of 3D scanning, allowing a realistic reproduction of the materiality of an artefact, and multispectral imaging, revealing hidden text. Interaction with the digital artefact is critical interpretation, and thus "digitization and visualization are [...] an integral part of the papyrological workflow,"[38] i.e., eventually, of the critical edition.

The digital object is not a mere, static copy of the original piece, but a dynamic component of papyrological scholarship, capable of reshaping the way in which we think the entire research process. Indeed, modern imaging techniques offer a more 'holistic' perspective on the material objects of study, as recently pointed out by Kathryn Piquette: while the texts are usually detached from their material support during their study, producing and handling a digital image enhance and stress not only the relationship between these two components, but also how the text was actually enjoyed in its ancient environment – the interactions between the material object and its users.[39]

Although this may seem a slight paradox – Papyrology gets more and more interested in papyri as material artefacts while Digital Papyrology deals with computerized information about papyri, which produces a dematerialization of the objects themselves – the increasingly close connection between text and image in the digital realm was already envisaged by Traianos Gagos, who – as early as 1998 – noted: "In this new era of papyrological research, we cannot speak of a collection of papyri alone, but also of a collection of electronic files, data, metadata and digital images."[40]

Here, the three key terms are: (1) data, i.e., the very texts stored in the databases; (2) metadata, i.e., the contextual information about texts (inventory, chronology, provenance, typology...), stored in the electronic catalogues, which themselves are part of the materiality of the papyri;[41] (3) digital images. The notion of 'meta-text', as introduced by Gagos, points to a digital interconnection among these components, in order to create an enhanced representation of the papyrus, complete of all its material identifiers:[42]

37 Tarte 2011a/b/c: 2012; 2016; see also Terras 2005 and 2006, 27–83.

38 Tarte 2011c, 13.

39 Piquette 2018, 111–5.

40 Gagos 2001, 516.

41 On the notion of 'data' and 'metadata' in Digital Papyrology see Reggiani 2017, 8–9, and 2019a, 17.

42 Lamé 2014 has described this idea (with reference to ancient epigraphs) through Foucault's philosophical concept of 'dispositive': the message of the text-bearing object can be completely understood in relation with a complex network of many other heterogeneous pieces of information. The ultimate purpose is "to digitize also the network that connected those information systems, instead of digitizing each individually".

> The availability of huge amounts of information in fully searchable textual form with accompanying images through these new media is altering drastically the definition of what constitutes a 'text', the way we experience reading it and, ultimately, the plurality of messages a text can offer to one or more readers. The new methods of presenting text with marked up images and the simultaneous availability of a variety of other research tools within the same electronic environment give us new ways of visualizing and approaching a given text. An edited text is no more a static, isolated object, but a growing and changeable amalgam: the image allows the user to look critically at the 'established' text and to challenge continuously the authoritative readings and interpretation of its first or subsequent editors.[43]

This perspective has very much to do with textual transmission: as in the ancient times the texts circulated in an inseparable whole together with their material support immersed in its context, now we have the chance to reconnect the very text of the digital edition back to its material frame, in an enhanced way, in order not to lose an important part of the transmission process. As seen in section 1, the main purpose of a digital critical edition must be the critical representation of the ancient text in all its constituent features, and any digital reproduction of its material aspect is just one of these features.

The digital connection between the written text and its material support can be realized in various ways. The easiest and simplest is enriching the text with a hyperlink pointing to an external picture (so, e.g., Trismegistos and several cases in Papyri.info – Fig. 5) or an external picture embedded in the same window as the text (so, e.g., HGV and Papyri.info when applicable – Fig. 6). The hyperlink is usually provided as part of a set of metadata that surround the papyrus text.[44] In the case of the APIS images embedded in Papyri.info records, the picture can be scaled. Of course, both strategies depend on the actual existence and availability of one or more digital pictures of the relevant papyri.

A tighter connection between text and image has been attempted in the framework of the *Anagnosis* project at the University of Würzburg.[45] On the one hand, hyperlinks to digital images of Herculaneum papyri (mostly consisting of early drawings and engravings) have been added to the encoded text in Papyri.info at the `<div>` level, i.e., in correspondence of each textual division, usually each column of the roll, rather than generally juxtaposed as in the previously mentioned cases (Fig. 7).[46] On the other hand, the project has developed a web-based tool to automatically align digital transcriptions of literary papyri with their corresponding digital images, in order to simplify various palaeographical tasks (handwriting analysis, gap filling, etc.).[47]

43 Gagos 2001, 514–5.

44 The digital text is native in Papyri.info, which is the source of both HGV and Trismegistos transcriptions (in the latter, not in all cases, and not for the literary and paraliterary papyri).

45 See Reggiani 2017, 154–6; Ast – Essler 2018, 68–73; Reggiani 2019a, 227–8; and the chapter authored by Vincenzo Damiani in this volume.

46 Other, non-Herculaneum literary papyri have been encoded like that in the framework of the same project.

47 For further details and screenshots see V. Damiani's contribution in this volume.

Digital alignment is currently the best option to represent the connection between the editorially transcribed ('absolute') text and the material written support in a virtual environment. Further developments in this direction have been accomplished by the D-Scribes project at the University of Basel, with the purpose of creating large datasets of character images that can train automated processes of handwriting recognition and related computer-assisted analyses.[48]

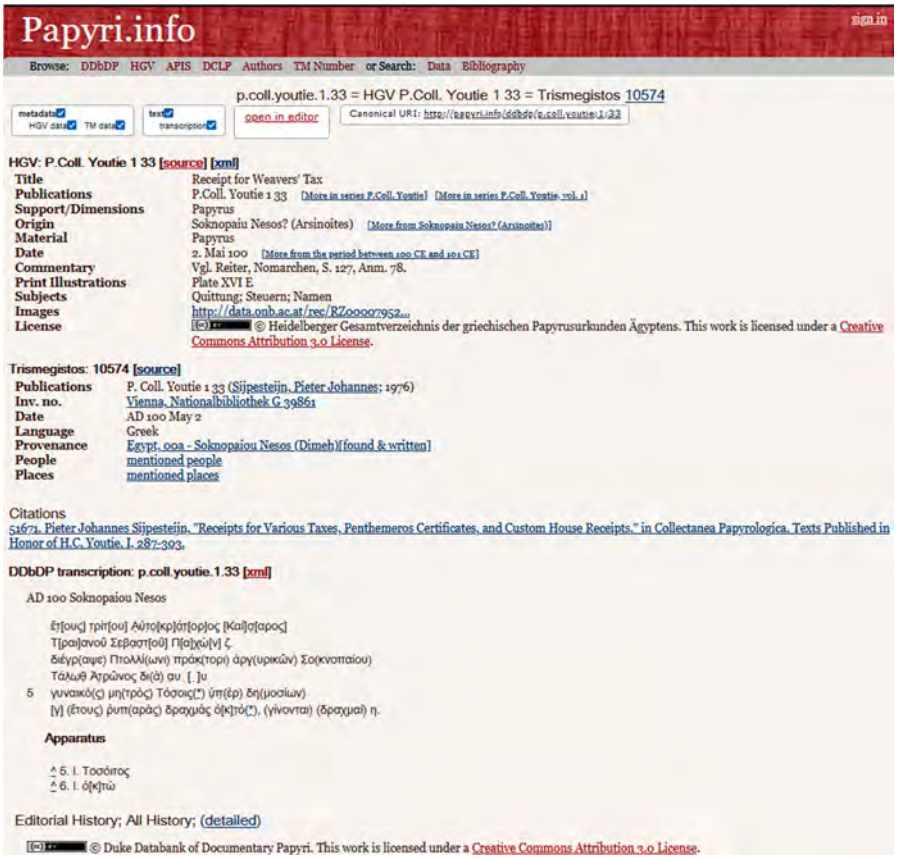


Fig. 5: Digital image hyperlinked from an external catalogue in Papyri.info (tenth field in the HGV metadata section) (https://papyri.info/ddbdp/p.coll.youtie;1;33).

48 For further details see the chapters authored by Isabelle Marthot-Santaniello and Olga Serbaeva, and by Nicole Dalia Cilia *et alii*, respectively, in this volume.

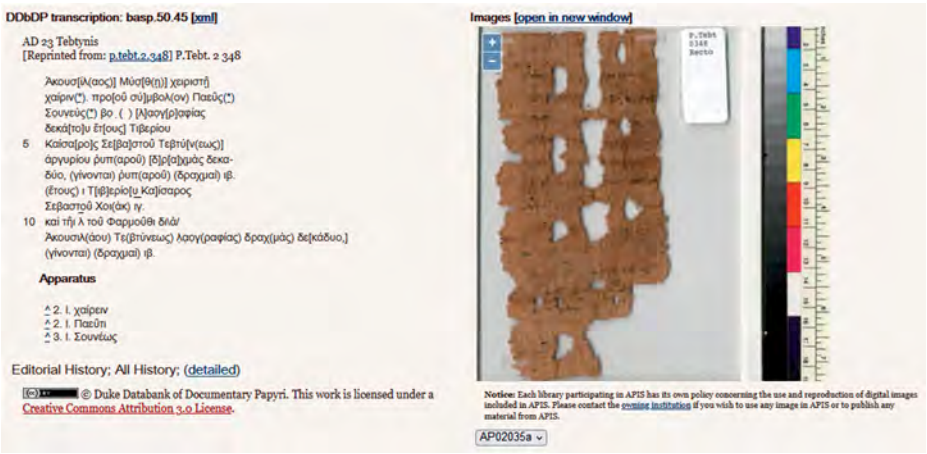


Fig. 6: Digital image embedded from the APIS catalogue in Papyri.info (https://papyri.info/ddbdp/basp;50;45).

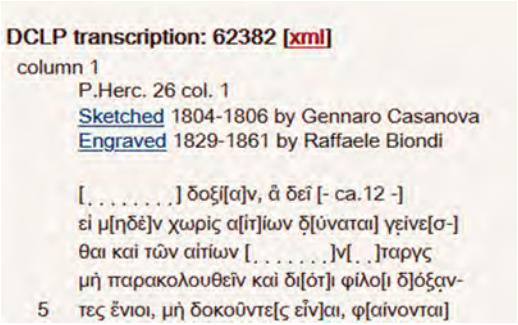


Fig. 7: Digital images hyperlinked from external resources in Papyri.info at the column level (P.Herc. 26 [TM 62382], https://papyri.info/dclp/62382). The underlying XML code is `<div n="1" subtype="column" type="textpart" corresp="#FR1340">`, where the `corresp` attribute points to the available images.

Such developments fall outside our current focus on the digital critical edition, but lead us to the second point at stake in this section: the materiality of writing, i.e., the writing process itself and the writing strategies deployed by the ancient scribes to make their texts capable to transmit the intended message. More or less automatic handwriting recognition is based on the same workflow that modern editors follow to decipher papyrus texts, and that ancient readers followed to understand the same texts: recognizing and decoding the shapes of the characters, their arrangement on the writing surface, the way of tracing the sequences of characters and words (the *ductus*), and so on. Digital alignment represents this workflow in both directions: (1) the extraction of the 'absolute' text from its written manifestation; and (2) the reunification of the 'absolute' text with its material counterpart (Figs. 8–9).

Fig. 8: This reading exercise offered by the University of Michigan Papyrus Collection (P.Mich. inv. 3196, line 1, https://apps.lib.umich.edu/reading/Zenon/line01.html) helps giving a rough visual idea of the workflow of reading/deciphering the original handwritten text and representing it in an 'absolute' virtual text. The editorial (interpretive) representation of the text would be (ἔτους) 31 μηνὸς Ἀθὺρ 12 ἐν Φιλα|δελφείαι κτλ., which more or less corresponds to the live understanding of the text by its ancient addressee.

Fig. 9: This sample palaeographical dataset of a typical papyrus from the Zenon archive (see Fig. 8) gives a rough visual idea of the correspondence between materiality (actual written characters) and abstraction ('absolute' transcription) in the digital palaeography of the papyri (https://apps.lib.umich.edu/reading/Zenon/paleography.html).

Furthermore, the material appearance of the text on (and in relation with) its support is given by the set of layout and graphical devices employed by the scribe to communicate with the reader. This has been scarcely appreciated in the early stages of Digital Papyrology, as long as it was considered a secondary feature of documentary papyri. Writing strategies like line displacements (*ekthesis*, *eisthesis*, centred titles), column alignments, abbreviation types were desultorily marked in the printed editions, and – consequently – largely neglected by the digital textual databases. A deeper interest in paratextual features has recently arisen thanks to the encoding of the literary and paraliterary papyri (*Digital Corpus of Literary Papyri*, DCLP), which make a massive and significant use of signs and layout to convey special meanings (reading help, critical marks, technical knowledge).[49]

Let us take, for example, the tax receipt preserved by P.Coll.Youtie I 33 (TM 10574, Soknopaiou Nesos, AD 100 – Fig. 10), which is always a good example because it is the papyrus chosen by William Willis to present the freshly launched Duke Data Bank of Documentary Papyri at the 17th International Congress of Papyrology (Naples, 19–26 May

49 In general, on the DCLP project see Ast – Essler 2018, 63–7. On the encoding of papyrus paratext see Reggiani 2017, 251–4; 2018a, 30–5; 2019a, 330–4; 2019c, 844–7; 2020.

1983),[50] thus allowing for historical comparisons with the current papyrological databases.[51] The editorial transcription of the text reads:

> ἔτ̣[ους] τ̣ρίτ[ου] Ἀ̣ὐ̣τ̣ο̣[κρ]ά̣τ̣[ορ]ος [Καί]ς̣[αρος] | Τ[ραι]ανοῦ Σ̣ε̣β̣α̣ς̣τ̣[οῦ] Π[α]χ̣ὼ̣[ν] ζ. | διέγρ(αψε) Πτολλί(ωνι) πράκ̣(τορι) ἀργ(υρικῶν) Ϙο(κνοπαίου) | Τάλ̣ῶ̣θ̣ Ἀ̣τ̣ρ̣ῶνος δι(ὰ) α̣υ̣ . [.]υ |[5] γυναικὸ̣(ς) μη(τρὸς) Τόϙοις ὑ̣π̣(ὲρ) δη(μοσίων) | [γ] (ἔτους) ῥυπ(αρὰς) δραχμὰ̣ς ὀ[κ]τ̣ό, (γίνονται) (δραχμαὶ) η.
>
> Apparatus: 5. l. Τοσόιτος | 6. l. ὀ[κ]τώ.
>
> Year third of Emperor Caesar Traianus Pius, Pachon 7. Taloth son of Atron, through [...] his wife, whose mother is Tosois, has paid to Ptollion, money-tax collector of Sonkopaiou (Nesos), for the public (taxes) of the year [3], eight impure drachmas, the total being 8 drs.

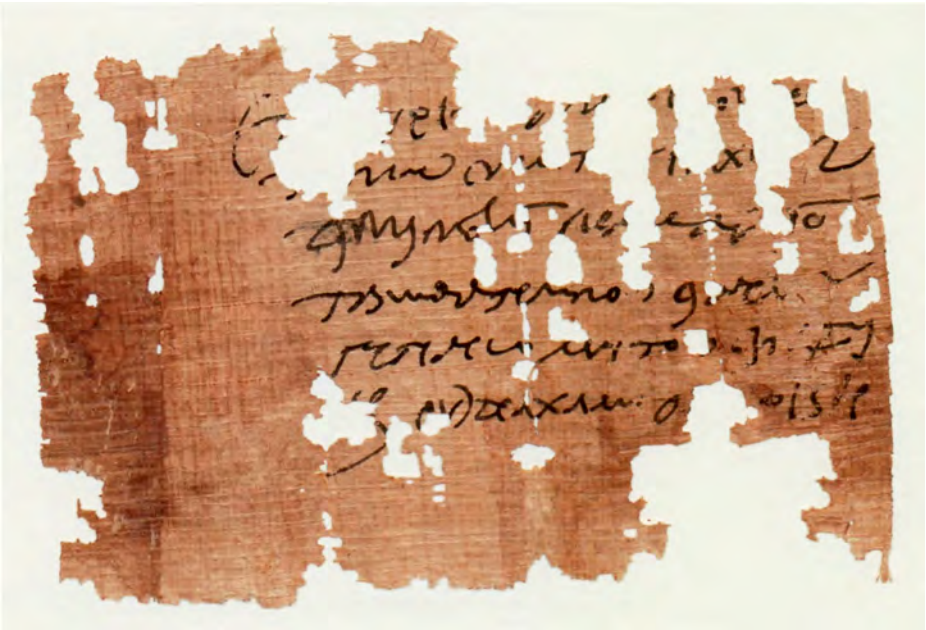


Fig. 10: P.Coll.Youtie I 33 (P.Vindob. inv. G 39861). © Österreichische Nationalbibliothek, Papyrussammlung.

It is a very simple and common text, yet exhibits several editorial peculiarities that are most interesting from the viewpoint of Digital Papyrology. For example, it clearly shows ancient writing strategies that have not been encoded in the original database (Fig. 2) and still lack from Papyri.info (Fig. 4): the most evident are the slight *ekthesis* of the first line, with an enlarged initial *epsilon* that marks the beginning of the document, and the

50 Willis 1984 (see also above, §1).

51 See above, §1, as regards the historical evolution of the digital treatment of editorial 'regularizations'.

quasi-monogrammatic rendering of the sequence *delta-iota* in the keyword διέγραψε at the beginning of the second line and in the preposition διά in line 4.

This is not a shortcoming of the digital encoding, since they lack from the source printed edition (Fig. 1) as well, and evidently, they were not considered relevant when the first database was conceived. It is also clear that a global reconsideration of the papyrus text from the material viewpoint can lead to further interesting editorial thoughts: for example, is the *delta-iota* in l. 4 really an abbreviation as printed in the edition and subsequently encoded in the database – i.e., δι(ά) –, or rather an elided δι' before the following *alpha*? Moreover, it seems that at the beginning of l. 3 the verb διέγρ(αψε) covers a previous writing, deleted by the ancient scribe. Since the outline of the cancelled text looks like the sequence *tau-alpha-lambda* that we find at the beginning of the next line, one may wonder whether this receipt was a copy of an original text and the scribe just started transcribing the wrong line before realizing the mistake. One may also notice that the surviving traces at the beginning of the last line are not entirely compatible with the supplement provided in the edition – i.e., [γ] (ἔτους) – and may well conceal β (ἔτους), which could mean that the payment was made for the previous year.

The different types of abbreviation employed in this papyrus are also significant of the multifaceted possibilities of visual interpretation and understanding of the writing strategies of ancient scribes. We have examples of symbols (l. 6: S-shaped ἔτους; slanting stroke for γίνονται; sinusoid for δραχμαί), truncations with special formatting of the last letter (l. 3: διεγρ with backward *rho*; πρακ with prolonged *kappa*; l. 5: δη with *ēta* attached to *delta*; l. 6: ρυπ with inverse rounded *pi*), plain truncations (l. 5: μη), overwritten letters (l. 5: υ$^{\pi}$ with rounded *pi* over *hupsilon*), overstrokes (l. 3: αρ$\bar{\gamma}$, ϙ$\bar{\text{ο}}$ that may well be quick renderings of the last letters as well, i.e., αργ$^{\upsilon}$, ϙο$^{\kappa}$). A deeper consideration of this topic leads to reflect on the abbreviation at l. 3 Πτολλί(ωνι), which looks as if the scribe started writing πτο$^{\lambda}$ (with *lambda* over *omikron*) and then, perhaps judging it too general, continued with λι at the line level, concluding with an overstroke that may well be a flat *omega* – in this case, it would be resolved as Πτολλίω(νι). Some of these abbreviations may have belonged to administrative practices, some to the scribe's own attitude. It is important to consider them in a global representation of the papyrus text as a writing phenomenon. Nevertheless, as is known, Papyri.info currently supports a general encoding for all type of abbreviations and symbols, without typological distinctions.[52]

The writing features, the layout strategies, and the paratextual framework deployed by the ancient scribes to convey the textual message is strictly related to the functional use and the actual circulation of the text itself. In turn, their technical use produced textual phenomena, the description and transmission of which go far beyond a traditional, 'static' philological model. Digital encoding provides a momentous opportunity to develop an accurate reproduction of the ancient texts presenting all their constitutive parts in an enhanced way. Moreover, it is an occasion to rethink and re-read

52 See Reggiani 2018a, 24–6; 2019a, Appendix, 21–2; 2019c, 846–7.

these texts, and to pay specific attention to peculiar features sometimes neglected or misunderstood by printed editions, devoted as they were – and sometimes still are – to the reconstruction of an 'ideal' text.

3 Ancient literacies and the digital edition of the papyri

> Die ersten vierzig Jahre unseres Lebens liefern den Text, die folgenden dreißig den Kommentar dazu.
>
> Arthur Schopenhauer[53]

In her seminal article about "Doctors' literacy and papyri of medical content," Ann E. Hanson showed how the multifarious evidence provided by the Greek papyri from Egypt bears witness to a widespread medical literacy in the Graeco-Roman world.[54] This term – "medical literacy" – is used to encompass a wide range of personalities, comprising both specialized physicians and learned laymen with specific interests in medicine and related topics, and points to the ability of reading, understanding and producing a written text dealing with technical subjects. In a subsequent contribution of mine, I tried to develop her overview, showing that ancient medical writings on papyrus are in fact characterized by multiple 'literacies' that can be better understood through the categories of (trans)textuality and better represented and studied in their complexity within a digital infrastructure.[55]

From a broader sociological point of view, the (plural) concept of 'literacies' has been developed to refer to "text-oriented events embedded in particular sociocultural contexts,"[56] stressing for example the use of reading/writing abilities, as well as communication strategies. This fits particularly well the situation of the papyrological sources, which show a complex degree of transtextuality that can be described through the models elaborated by Gérard Genette since the Eighties. Transtextuality defines all the various possible relationships among texts – "all that sets the text in relationship, whether obvious or concealed, with other texts"[57] – and encompasses five subcategories that are sketched as follows:[58] (1) *Intertextuality* is the relation between parallel text, in the form e.g. of quotation or allusion; (2) *Paratextuality* is the relation between one text and what

53 ["The first forty years of our life provide the text, the following thirty the commentary."]
54 Hanson 2010.
55 Reggiani 2019d.
56 Johnson 2009, 3.
57 Genette 1992, 83; then Genette 1997, 1.
58 Based on Genette 1992, 83–4, as later developed in Genette 1997, 1–7.

surrounds the main body of the text (e.g. titles, headings, graphical/layout devices);[59] (3) *Metatextuality* is the explicit or implicit critical commentary of a text on another text; (4) *Hypotextuality*/*hypertextuality* is the relation between a text and a preceding hypotext that is transformed, modified, elaborated or extended; (5) *Architextuality* is the designation of a text as a part of a genre or genres.

The bottom line is that a text was not transmitted alone, as an independent and isolated message, within the material frame discussed above in section 2, but as a part of a complex network of cultural cross-references. The most logical examples are in literature and – even more – in technical works, where, among influences, borrowings, direct or indirect quotations, allusions, annotations, commentaries, each one of the abovementioned stages is deeply interrelated with the others.

A particularly relevant case is offered by the so-called *Michigan Medical Codex* (P.Mich. XVII 758 [TM 59332], 4th century AD), a collection of medical recipes commissioned by a practicing physician:

> First he collated the text of his newly-made copy against an exemplar, making corrections in addition to the items already corrected by the scribe, and then he went on to more than double the contents of the codex by filling the margins with additional recipes for pills to medicate bodily ills and plasters to medicate wounds and lesions of every kind. Naming a therapeutic recipe after the physician or pharmacologist from whose works it had been taken, or by whom it was popularized, became increasingly common in Hellenistic and Roman times, and the codex cited recipes attributed to a number of medical authors [...]. The recipes in the codex frequently show correspondences with recipes for plasters in the collections of Galen, Oribasius, Aëtius, or Paul of Aegina that have come down in the manuscript traditions, highlighting the striking degree of continuity among ingredients and their relative proportions from hand-written copy to hand-written copy over many centuries.[60]

Intertextuality, hypotextuality and similar connections merge together, creating a very complex and unique clockwork: "although individual recipes in a collection on papyrus often resemble items in the known authors, each extensive collection on papyrus has thus far proved to be a unique assemblage."[61] The paratextual function of critical and lectional marks stresses the 'composite' structure of the text, acting as a bridge between its original oral roots and its later outcome.

The inadequacy of the traditional philological/stemmatological model to represent in full the compositional stages and the textual features of these complex and fluid technical writings has already been pointed out by Ann Hanson herself in an earlier work.[62] This corresponds to the inadequacy of the traditional 'literacy' model to describe and

59 In Genette's view, paratext is mainly related to titles, headings, and other textual elements that surround the main text. For a broader view of 'paratext' including graphical marks and layout arrangements, see Reggiani 2017, 202–3; 2018a, 30; 2019a, 250–1; 2020, 184 n. 15; Choat 2021; Reggiani 2023a, 134–5.

60 Hanson 2010, 197–8.

61 Hanson 2010, 199.

62 Hanson 1997.

understand the textual facts. The "accretive model of composition," advanced by Hanson to provide a suitable description of the phenomenon, seems to me perfectly pertinent to the plurality or network of 'literacies' revealed by the texts in question. The witness is unique, but the transtextual relationships create a multifarious network that clearly goes beyond the mere fixation of a canonical archetype and does justice to a complex and fluid interconnection of multiple literacies.[63]

Quotations, allusions, reuses are of course a daily matter in proper literature,[64] so that it is not worth spending much time to discuss them – I would rather focus on different typologies of text use, reuse, and transmission, which are peculiar of the papyrological sources. First, all the formulaic phrases that are so commonly used in both private (letters, accounts…) and public (petitions, reports, receipts…) documents from Graeco-Roman Egypt.[65] They are recurring wordings used as standard identifiers of specific sections of the said documents and they belonged to the scribes' and to the wider public's collective literacy.[66] There is no direct derivative relationship among the same type of documents, but they belong to well-established textual schemes, which are transmitted, more or less unchanged, over the time. Second, the peculiar form of 'quotation' that is represented by the insertion in a document of 'copies' (ἀντίγραφα) of other documents. This was a particular communication strategy employed for the exchange of information in various sectors of ancient life. We can find this quotation pattern in several different document typologies, for example official letters and other public documents that can embed many other letters, orders, laws quoted in a cascade pattern that has been successfully analyzed (for the Ptolemaic period) by Giuditta Mirizio.[67] Third, literary echoes in documentary texts, in terms of either direct quotations or more or less indirect allusions.[68]

All the preceding intertextual manifestations are not confined to the pure or absolute 'text', but do intertwine with the materiality of the writing support in terms of palaeographical format and paging layout. Formulaic patterns in letters often feature

63 See also Reggiani 2019e and 2023a.

64 Among many possible papyrological examples, see e.g. Eckerman 2010, with the edition of a late-antique papyrus containing an hexametric encomium with an allusion to Hom. *Il.* II 489.

65 Some documentary and paraliterary genres have received systematic attention as regards formulaic and recurring patterns: e.g. private letters (Evans 2007; Luiselli 2008, 692–707; Porter – Pitts 2013; Nachtergaele 2013 and 2016), petitions (Ptolemaic: Baetens 2020; Roman: Mascellari 2021; Byzantine: Fournet 2004); reports of trial proceedings (Coles 1966); medical prescriptions (Gazza 1955; Andorlini 2017, 15–36); medical reports (Reggiani 2018c); iatromagical recipes (de Haro Sanchez 2015); legal documents (Laffi 2013).

66 On the relevance of formularies in the scribes' professional training and practice see Migliardi Zingale 2003; Bucking 2007.

67 See Mirizio 2016 and 2021.

68 In general, on poeticisms in documentary papyri, with particular discussion of Dioscorus' paperwork, see Luiselli 1999, 189–213. On Homeric quotations and echoes in documentary papyri, again with a particular focus on Dioscorus of Aphrodito, see Fournet 2012. On New Testament references in papyrus letters see Choat 2006.

specific visual strategies aimed at highlighting their special communicative purposes[69] just as recurring textual structures in prescriptive technical texts employ equally recurring formatting architectures.[70] Quotations of any kind – either documentary or literary – can be often emphasized by means of graphical or layout stratagems.[71] Generally speaking, it has been recognized that most of the papyrus texts comply to what has been defined a "meaningful palaeography", that is the deployment of specific palaeographical or layout strategies to underline the meaning of the text, its purpose, its context.[72] This is a cognitive textual network that operates at two different but intersected levels: that of transtextuality, which we are dealing with in this section, and that of the materiality of writing, discussed previously in section 2. It must be noted that such an interplay works at every degree of transtextuality as defined by Genette's scheme introduced above: for instance, not only intertextuality (formulas, quotations, allusions) and paratextuality strongly influence – and are influenced by – palaeography, but also architextuality, in that the 'genre' of a text – either literary or documentary – often relies on specific material formats.[73]

A single, significant example may suffice. P.Flor. II 259 (TM 11146, Theadelphia, AD 249–268 – Fig. 11) is a letter from Timaios urging Heroninos to complete a payment in grain. The editorial transcription of the main text reads:

> Τίμαιος Ἡρωνίνωι τῶ[ι] | φιλ(τάτῳ) χαίρειν. | κἂν νῦν καιρὸν ἔχεις ἀ|ναπέμψαι ἢ τὰ σιτάρια |[5] ἢ τὴν τιμὴν καὶ μα|θέτω ὁ Κιοτ' ὅτι ἐὰν μὴ | δη τὸν ἄλλον σάκκον ἢ ἀ|νέλθῃ καὶ τὸ κατ' αὐτὸν ἔν|θη στρατιώτης κατέρχετε |[10] ἐπ' αὐτόν. ἀλλὰ πάντως ἀ|ναπέμψον αὐτὰ. ἐρρῶσθαί | σε εὔχομαι.
>
> Apparatus: 3. l. καὶ ἂν | 6. κιοτ' pap. | 7. l. δῷ | 9. l. κατέρχεται.
>
> Timaios to the beloved Heroninos, greetings. It is now the right time to send me either the sacks of grain or the (corresponding) price. Let Kiot know that if he doesn't deliver the other sack or come here and pay what he owes, a soldier will come to collect it from him. But you must send them to me, absolutely. I hope you are well.

The same hand – well recognizable from the pen trait, the size of the characters, the writing grade, the rendering of some ligatures – employs three different writing styles according to the sections of the text.[74] The main body of the letter (ll. 1–11) is penned in a semilibrary writing close to the 'severe style' found in many contemporary literary papyri like the *Hellenica Oxyrhynchia* (PSI XIII 1304 [TM 59482]). However, the first two

69 See Sarri 2018, 114–92.

70 See de Haro Sanchez 2015; Reggiani 2022b; Monte 2024; Bongiovanni 2024.

71 See Mirizio 2021, passim.

72 See Fournet 2007 and 2022.

73 The *grammateus* project at the University of Geneva is precisely devoted to the digital categorization of typologies of papyrus texts according to the structure, format, and layout of the documents: see the chapter authored by Elisa Nury in this volume.

74 See Messeri 1998, 208; Reggiani 2019f, 177–9.

lines, comprising the initial address and greetings formula, are graphically separated from the rest of the body, emphasized by means of slightly wider interlinear spaces, slight *eisthesis* of both lines, and blank spaces between the words. The final greetings (ll. 10–11) are traced quicklier: the handwriting becomes smaller and cursive, acting as the sender's 'signature'; again, they are emphasized by means of a special layout (the first verb, ἐρρῶcθαι, fills in the remaining space of the last line of the letter, while the rest of the closing formula is written starting from the middle of the writing column). The textual content is not an absolute communication medium, but is enriched by a complex material context that allows the writer to send the reader further information.

Fig. 11: P.Flor. II 259. Florence, Biblioteca Medicea Laurenziana. By permission of the MiC. Any further reproduction is forbidden by any means.

Moreover, an addition in the left-hand margin – perpendicular with respect to the main text – quotes the first two verses of the second book of Homer's *Iliad*: ἄλλοι μὲν ῥα θεοί τε καὶ ἀνέρες ἱπποκορυσταὶ | εὗδον παννύχιοι, Δία δ' οὐκ ἔχε ν[[υ]]ήδυμος ὕπνος | εὗδον παννύχι⟨οι⟩.[75] This marginal quotation was penned in an upright library hand by the same Timaios and acts as an erudite, sarcastic joke referred to the current situation of delayed payment complained in the letter.[76] The connection between the literary quotation and the letter is assured not only by palaeography (same hand) and layout (the use of the left-hand margin, commonly hosting *post scriptum* additions to letters in Roman times[77]), but also by the repetition of εὗδον παννύχι⟨οι⟩ that emphasizes and discloses the allusion. In this case we are facing a double transtextual relationship: intertextuality and metatextuality at the same time.

A different opinion formulated by Colin Roberts, that "an unknown hand has added the Homeric tag in the margin in a conscious approximation to a literary hand,"[78] seems to have influenced the alternative idea that the Homeric quotation had nothing to do with the letter. This unlikely possibility has nonetheless induced the digital catalogues to consider the possibility that the added hexameters represented a school exercise independent from Timaios' reprimand: Trismegistos devotes a specific record to the quotation (TM ID 60203 = LDAB ID 1320 – Fig. 12) considering it a "reuse of blank space" of the letter (TM ID 11146);[79] Mertens-Pack[3] recorded it as "ex(ercice) d'écriture?" (MP[3] ID 623 – Fig. 13).[80] Accordingly, the Homeric quotation received an autonomous DCLP record on Papyri.info (Fig. 14),[81] while it was already added to the DDbDP record of P.Flor. II 259 (Fig. 15),[82] thus generating an amusing short circuit that illustrates the difficulties of current digital platforms to handle such complex textual interactions.

75 "The other gods and the chariot-charging men slept the whole night through, but sweet sleep came not on Zeus" (transl. Powell 2014).

76 So D. Comparetti *ap. ed.pr.* (1908); Rathbone 1991, 12–13; Cribiore 1996, 6–7; Messeri 1998, 208; Cribiore 2001, 179; Fournet 2012, 141–2; Del Corso 2018, 37–8; Fournet 2018, 193; Larsen 2018, 477; Reggiani 2019f, 178. Collins 2008, 227–8 advances the hypothesis that the literary quotation should be interpreted "in light of the magical and divinatory use of Homeric verses," so that "Timaeus put the Homeric verses there to ensure that in the larger scheme of things, his goods would be protected, no matter what Heroninus did." But the content of the letter depicts a slightly different situation and there is no other hint that the verses were intended to be something more than an erudite reproach. Collins' view is judged "peu vraisemblable" also by Fournet 2012, 142 n. 72.

77 See Luiselli 2008, 707–9; Sarri 2018, 112–3.

78 Roberts 1956, 22.

79 The field "Content" registers: "school text? or postscriptum to a letter (private)?" (https://www.trismegistos.org/text/60203).

80 The annotation appeared in the previous versions of the catalogue (cited also by Fournet 2012, 142 n. 66); it has now disappeared from the most recent version (http://www.cedopalmp3.uliege.be/cdp_MP3_ display.aspx?numnot=00623.000).

81 https://papyri.info/dclp/60203. The text was encoded by M. Sampson in 2023.

82 https://papyri.info/ddbdp/p.flor;2;259. The text of the letter (without the quotation) was transcoded from the old Duke Databank of Documentary Papyri in 2008; the text of the quotation was added by R. Ast in 2015 as a marginal addition attributed to a second hand.

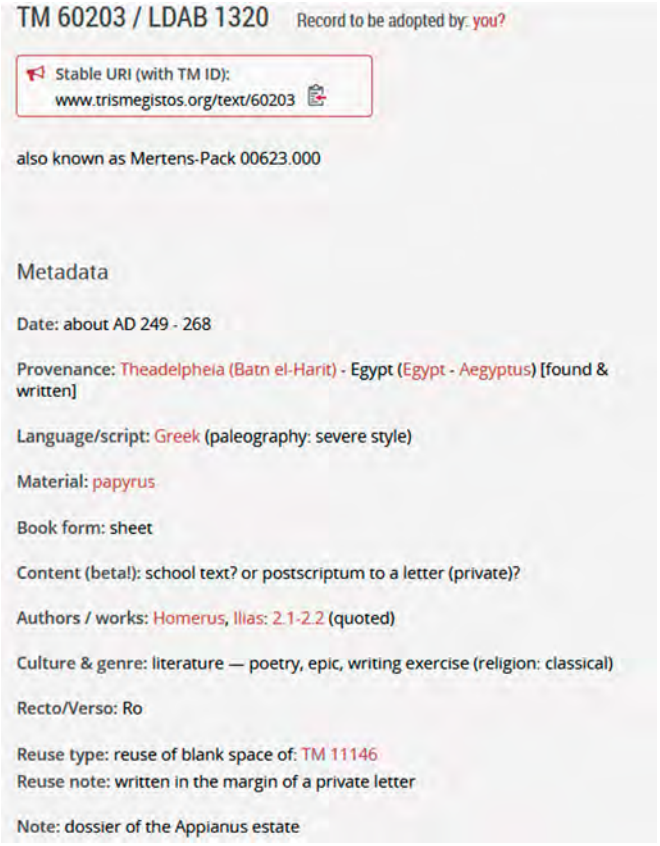


Fig. 12: Trismegistos record of the Homeric quotation in P.Flor. II 259.

Fig. 13: MP^3 record of the Homeric quotation in P.Flor. II 259 (previous versions).

Introduction

This text is written in the left margin, perpendicular to the text of a letter from Timaios to Heroninos (P.Flor. 2.259).

DCLP transcription: 60203 [xml]

ἄλλοι μέν ῥα θεοί τε καὶ ἀνέρες ἱπποκορυσταὶ
εὗδον παννύχιοι, Διὶ δ' οὐκ ἔχε νήδυμος ὕπνος,
εὗδον παννύχι

Editorial History; All History; (detailed)
2023-01-05T13:25:52-05:00 [Mike%20Sampson]: Finalized - This is good to go.
2023-01-05T12:16:09-05:00 [Mike%20Sampson]: Vote - Accept-Straight-to-Finalization - Yes.
2023-01-05T12:13:42-05:00 [Mike%20Sampson]: Submit - Added text, from ed. pr. Please update metadata during finalization.

Fig. 14: DCLP record (with editorial history) of the Homeric quotation in P.Flor. II 259.

DDbDP transcription: p.flor.2.259 [xml]

III spc Theadelphia

ctr

Τίμαιος Ἡρωνίνωι τῶ[ι]
φιλ(τάτῳ) χαίρειν.
κἂν(*) νῦν καιρὸν ἔχεις ἀ-
ναπέμψαι ἢ τὰ σπάρια
ἢ τὴν τιμὴν καὶ μα-
θέτω ὁ Κιοτ' ὅτι ἐὰν μὴ
δη(*) τὸν ἄλλον σάκκον ἢ ἀ-
νέλθῃ καὶ τὸ κατ' αὐτὸν ἐν-
θη στραπώτης κατέρχετε(*)
ἐπ' αὐτόν. ἀλλὰ πάντως ἀ-
ναπέμψον αὐτά. ἐρρῶσθαί
σε εὔχομαι

ms

(perpendicular) (hand 2) *(Iliad 2.1-2)* ' ἄλλοι μέν ῥα θεοί τε καὶ ἀνέρες ἱπποκορυσταὶ
(perpendicular) εὗδον παννύχιοι, Διὶ δ' οὐκ ἔχε νήδυμος ὕπνος '
(perpendicular) εὗδον παννυχί

Apparatus

^ ctr.3. l. καὶ ἂν
^ ctr.6. κιοτ' papyrus
^ ctr.7. l. δῷ
^ ctr.9. l. κατέρχεται

Editorial History; All History; (detailed)
2016-06-23T09:09:13-04:00 [berkes.lajos]: Finalized - ready
2016-06-23T09:04:36-04:00 [berkes.lajos]: Vote - Accept-Straight-to-Finalization - Add quotation mark.
2015-11-16T16:40:48-05:00 [simoeis]: General - In line 8-9, change ἔνθῃ to ἐνθῇ (cf. Preisigke WB s.v.). Seems to have been a data entry error.
2015-11-16T15:00:13-05:00 [simoeis]: Submit - I've added the two Homeric lines recorded in the left margin of this letter. If, as is commonly believed, they were intended to accompany the letter, then they should be included with it here.

Fig. 15: DDbDP record (with editorial history) of P.Flor. II 259.

Note also that – just as the previously mentioned case of P.Coll.Youtie I 33 – Papyri.info does not render the layout structure of the letter, except for the marginal addition (Fig. 15). In this case, however, the reason does not lie only in the paper edition, which on the contrary prints a quasi-diplomatic transcription of the text, highlighting some of the visual strategies of the ancient writing (Fig. 16). Papyri.info simply inherits the early DDbDP idea of 'absolute' text and traditional critical edition.

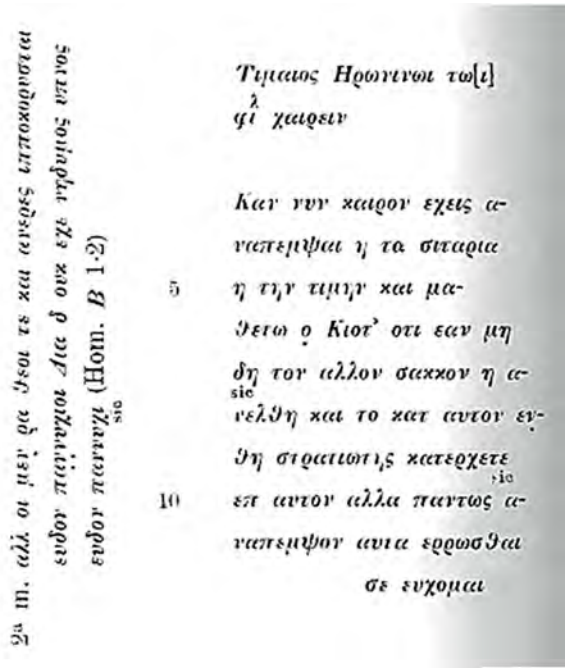


Fig. 16: *Editio princeps* of P.Flor. II 259.

The new digital tools allow us to reconsider the viewpoints discussed above and to develop new infrastructures in order to enhance the digital edition of such complex ancient sources, so that the final product is not a 'static' layout simply reproducing the paper editions, but something closer to the original sense of the text, to its network of literacies, and to its transmission. It is in particular in the possibility to develop multi-layer annotation schemes that we can find the most suitable architecture for representing complex and fluid textual products stemming from the horizon of the multiple 'literacies' sketched so far.[83] Creating a multi-layer edition means to give a digital representation of all the transtextual connections that make up the network of ancient literacies as described above, translating them into a new network of 'digital literacies', meaning the capability to handle multiple textual and cultural traditions at the same time. Of course, it is not possible to leave digital multi-layer editions of the literary manuscript tradition out of consideration, because of the thick intertextual links between the technical writings preserved on papyrus and the literary treatises, but also because of the interesting philological variants that appear in the papyrological sources.[84]

83 It is remarkable, from this point of view, the EVWRIT project and its multi-modal approach to various layers of papyrological data: see the chapter authored by Klaas Bentein in this volume.
84 See below, §4.

Again, this was envisaged by Traianos Gagos when, speaking about 'meta-papyri'[85], he wrote:

> Furthermore, the simultaneous access to and study of thousands of texts and their images that could be as far apart as a millennium, in a single search and through the same medium, has the potential to challenge our established notions of the "messages" a text carries within itself, its textuality and intertextuality [...]. As Roland Barth [*sic*] explains: "Any text is an intertext; other texts are present in it, at varying levels, in more or less recognizable forms: the texts of the previous and surrounding cultures. Any text is a new tissue of past citations. Bits of codes, formulae, rhythmic models, fragments of social languages, etc. pass into the text and are redistributed within it, for there is always language before and around the text". [...] In one or another way, papyrologists have always recognized the "intertextuality" of the Greek papyri from Egypt, because of the multicultural and multi-ethnic environment in which these texts were born. The development of the new electronic media in our field and the capability to establish these cross-links – or these intertextual signifiers, so to speak – on the linguistic, cultural and historical level through the interaction of multiple texts, images and a variety of related tools places the notions of textuality, intertextuality and metatextuality on a new (electronic) platform which, in turn, becomes part of these notions as the "carrier", "interpreter" and "distributor" of these texts.[86]

Note that also fragmentation, which is an intrinsic feature of the materiality of the papyri but also affects texts and their transmission to us, could be handled with the conceptual instruments of the theory of transtextuality. As it has been analysed by Monica Berti with reference to historiographical fragments,[87] transtextuality defines the various possible relations among texts, and we may refer it not only to a network of quotations and parallel passages ('intertextuality'), but also to the aspect of fragment which very often the papyrus, be it literary or documentary, shows. Indeed, we can define the fragmentary character of the papyrus as a sort of 'non-voluntary quotation', selected by the chance and by the material circumstances rather than by an author's will. The transtextual link will work, in this case – following the specific terminology –, as a relationship of 'hypertextuality', which describes the derivative connection of a 'hypertext' (in our case, the original document, lost, more or less recoverable) with a 'hypotext' (our fragment), showing various degrees of physical and mechanical interferences. As with the digital encoding of fragmentary quotations, "hypertextual models allow to rethink the fundamental question of the relation between the fragment and its context, representing and expressing every element of printed conventions in a more dynamic and interconnected way."[88]

The new possibilities offered by the digital infrastructures can pave the way to the creation of a larger and dynamic corpus of texts, where a multi-layer architecture can help giving ancient sources back their deep link with ancient literacies through new, modern digital literacies, in a continuous and fluid transmission flow.

85 See above, §2.
86 Gagos 2001, 516.
87 See Berti 2010.
88 Berti 2010, 1.

4 The digital concept of variant

> I have called this principle, by which each slight variation, if useful, is preserved, by the term of Natural Selection.
>
> Charles Darwin

The traditional idea of variation as a deviation of a text from its archetype throughout transmission over the time stresses the similarity between linguistic and philological variants:[89] both conceal the assumption that we need to emend a text in order to reach a virtual textual correctness, and in both cases the critical editor will print what (s)he assumes to be the 'correct' form in the text, relegating the deviating 'anomaly' in the apparatus. While a philological variant is usually defined after a comparison with other versions of the same text, papyri mostly appear to be unique texts ('single witnesses') and their 'variants' and 'errors' are usually intended as related not to an archetypical text, but to a standard reference language. One of the most striking editorial outcomes of the choice of this 'linguistic archetype' is the somehow fluctuating treatment of word forms that deviate from 'classical' Greek. This issue has been discussed in details elsewhere,[90] so I will just summarize the main points at stake.

As a tacit rule, Koine Greek forms, which are in fact 'linguistic variants' with respect to classical Greek, are commonly assumed to be the 'regular' forms in the papyri, in a more or less conscious consideration of the cultural and linguistic environment of Hellenistic and Roman Egypt. Nevertheless, the situation is not that clear and sometimes we do find Koine forms regularized with the classical standards. 'Irregular' word forms are irregularly dealt with by modern editors, with a not negligible outcome in terms of digital texts, in that the editors' original mood is usually retained, generating an evident loss of information in some cases.

Which begs the question: is the choice of a linguistic archetype effective for contexts in which linguistic changes occurred?[91] More generally, can the choice of a linguistic archetype be universally effective for a multilingual society? How much role does the frequency of attestation of a form play? Trevor Evans has demonstrated, through examples from the Ptolemaic archive of Zenon, the importance of considering terms of comparison among the papyri themselves in order to conceive a more or less correct idea of linguistic 'standard',[92] or better, in his own words, of "substandard usage in

89 This section develops a hitherto unpublished talk, titled "La digitalizzazione dei manoscritti antichi e l'edizione critica digitale", delivered at the conference "Parma 32°43'30° - I manoscritti greci della Biblioteca Palatina: codice, testo e immagine", Biblioteca Palatina, Parma, November 28, 2019. I warmly thank Massimo Magnani for inviting me to participate in that conference.

90 See Reggiani 2018a, 7–8 and 26–9; Stolk 2018; Reggiani 2019b.

91 See Stolk 2015a/b/c; Depauw – Stolk 2015.

92 See Evans 2010a/b and 2012a/b.

documents of the same place and time": "we should be building our understanding of an emerging standard language in non-literary papyri from this internal evidence much more than from the practices of classical literature",[93] as he convincingly concludes. Furthermore: how much role do personal consciousness and individual preferences or customary habits play?[94] In Sir Kenneth Dover's words, "[n]o utterance is such that its author cannot care what it sounds like:"[95] why should not we care it as well? Should we regularize according to our own linguistic taste or according to the ancient author's one? Note that the purpose of textual criticism is to establish what an author exactly wrote, and that, by definition, a linguistic variant is any of the different phonetic, morphological, or graphical aspects under which a word can appear in a language, the choice of which can be due to personal reason and preferences, or to archaic, regional, technical, poetic uses. It is a lively instance of text transmission in its actual stages, where every level of the papyrological witness is important to contextualize the text and, as a consequence, its critical interpretation:

> When we assess individual texts, we have to consider diachronic changes within the Greek language, linguistic register, educational levels, the circumstances and process of composition revealed by palaeography and format, even the difficulty of analysis derived simply from our all too frequent lack of contextual information. Without addressing these factors we cannot expect to achieve a satisfactory appreciation of the material.[96]

On the morpho-syntactic level, we witness phenomena that intertwine with other layers of communication. In medical prescriptions, it is remarkably frequent – not to say formulaic – the use of the verb χράομαι in the imperative form χρῷ "use" to introduce specific instructions about the administering of medicaments. This is typically accompanied by the indication of the excipient substance with which the remedy is to be taken[97] or of the ingredient to be used.[98] Nevertheless, it is not rarely the case that the syntagm "use with water" appears under the a-syntactic aspect ὕδωρ χρῷ,[99] which goes

93 Evans 2010b, 205

94 It is the case, for instance, of what C. C. Edgar called Amyntas' "weakness" for ἀφέσταλκα: the preference accorded by one of the main characters of the Zenon archive for the aspirated perfect form of ἀποστέλλω, instead of classical ἀπέσταλκα (P.Cair.Zen. I 59047, 1 n.; see Evans 2010a). This is certainly not a regular form, not even a correct one, but what to think when an author uses with constancy such an irregular form? Shouldn't we assume it as standard (or, according to Evans' terminology, "substandard"), since it was almost systematically (perhaps consciously?) employed by an author? And shouldn't we reverse the situation, positing the classical form as a variant of the idiosyncratic spelling?

95 Dover 1997, 24.

96 Evans 2012a, 123.

97 E.g.: SB VIII 9860 = GMP III 14 (TM 65669), ii 9 χρῶ ἐν ὕδατι "use in / with water"; P.Tebt. II 273 = GMP II 5 (TM 63789), ii 13 με]τ̣ ᾽οἴνου χρῶι "use with wine".

98 E.g.: P.Oxy. VIII 1088 (TM 63118), i 19 χαλκίτιδει λήᾳ χρῶι "use pounded rock-alum"; P.Oxy. LXXIV 4975 (TM 119320), fr. 1, 4 τῇ σποδῷ χρῶ "use the powder".

99 E.g.: P.Tebt. II 273 = GMP II 5 (TM 63789), iv 5, vii 17, viii 5, 22; P.Princ. III 155v (TM 63920), 9; P.Oxy. LXXIV 4977 (TM 119322), 1.

far beyond an apparent 'incorrect' anacoluthon, becoming a distinctive mark of medical recipes, sometimes further stressed with peculiar abbreviations (Fig. 17).[100] It would seem rather senseless to 'regularize' such peculiar circumstances, for which we may well speak of 'formulaic substandards', which increasingly tend to detach from the syntactic architecture of the discourse and to constitute textual and graphical units, completely released from the context. On the other hand, it would also be important to record the 'standard' linguistic form, so that a database query could effectively retrieve all the possible occurrences of the word or the phrase in the corpus.

Fig. 17: (a) ὕδωρ χρῶι in P.Tebt. II 273v, vi 28 (Courtesy of the Center for the Tebtunis Papyri, University of California, Berkeley); (b) ὕδωρ ⳨(ῶ) in P.Princ. III 155v, 9 (P.Princ. inv. AM 11224 B, photo provided in open access by the Princeton University Library's Papyri Collection); (c) ὕδ⸌ ⳨(ῶ) in P.Oxy. LXXIV 4977, 1 (courtesy of the Egypt Exploration Society and the University of Oxford Imaging Papyri Project); (d) ὕ^{δ}(ωρ?) ⳨(ῶ) in PSI X 1180, fr. B, iii 10 (Florence, Biblioteca Medicea Laurenziana, Ms. PSI 1180. By permission of the MiC. Any further reproduction is forbidden by any means).

Furthermore, let us consider the case of PSI X 1180 (TM 63458), fr. B. ii 10, a collection of recipes where the said formula occurs in a special abbreviation of both ὕδ- (*hupsilon* with *delta* above) and χρῷ (the usual *chi-rho* monogram) (Fig. 17d). The resolution of the abbreviation is usually printed as ὕδ(ωρ) based on the well-known syntagm, but nothing assures that it would not have been a 'regular' ὕδ(ατι), since we have no other instances of a-syntactic ὕδωρ in the extant fragments of this collection of recipes. The problem is the same as the long-standing and still unresolved issue of supplying material gaps potentially containing irregular linguistic forms. This was already pointed out by Greg Horsley at the 20th International Congress of Papyrology (Copenhagen 1992) in the following terms:

> Thanks to often easily identifiable types of texts, many very broken papyri and inscriptions can be restored convincingly by analogy with their generic cousins. The matter to raise here is, once more, a plea to the editor by the user concerned with the language. A restoration which is very likely correct is of unequal value according to the reader's interest. An historian may have sufficient detail from what survives for the restoration to serve to fill in the gaps in the text adequately; the lacuna may make no material difference to the ability to use that papyrus for historical argument. The situation may be rather different for the person investigating syntactic usage. For once a restoration is included which follows a 'normal' syntactic construction, a predisposition is created for the reader that this was the reading. Frequently, of course, the editor's judgement is very sanely based; but there are always 'wild cards', the unexpected lexical, syntactic, and morphological

100 See Reggiani 2022b, 125–8.

> usages which can be the 'tip of the iceberg' for the alert philologist. Is a lexicographer to claim a partly surviving word as an attestation of that word? [...] Sometimes, too, editors restore their papyrus texts according to the orthography of the surviving portion (itacism, Atticistic features, etc.); but this is not always the case, for sometimes the restorations are given in normalised Greek. [...] The risk is that this procedure creates a predisposition for the incautious reader to accept that the normalised form is what the ancient writer intended.[101]

The critical uncertainties are of course mirrored by the digital uncertainties: the impression is that the extant syntax to encode linguistic variants – which reflects the current scholarly position through the lens of the Text Encoding Initiative standards – is not really designed to represent complex cases of potential substandards, and this necessarily turns into simplifications that do not correspond to the real communicative network embedded in the papyrus texts. When we encode the text of a papyrus, we must take a decision. Encoding means indeed to transfer the text in a machine-readable language that is conventional, logical, precise and standardized. Any possible uncertainty may result in potential loss of information and therefore in limitations to the enormous potentialities of the database. For example, the current markup tag used to indicate a linguistic variation of any kind is called 'regularization' (`<reg>`) and involves a rigid separation between the 'original' (`<orig>`) reading (displayed in the main text) and the 'regularized' spelling (displayed in the apparatus). It is evident that behind such a syntax lies the traditional idea that any variant must be brought back to a form that is assumed as regular. This is understandably affected by the uncertainty and inconsistency in defining a 'standard' form, which is affected – in turn – by discussion about the very nature of linguistic variation.

The switch from the printed medium, intrinsically limited, to the digital space, which offers potentially endless possibilities of handling the texts, is a momentous occasion for rethinking the concept itself of textual variation (of any kind). A significant improvement in the addition of meaningful information comes from linguistic annotation, a powerful methodology developed by corpus linguistics, the branch of linguistic studies that deals with corpora of texts as representative samples of an entire language. Annotating a corpus means to tag textual elements in a systematic way, adding some kind of linguistic information.[102] It allows describing, recording, understanding, interpreting, and analyzing linguistic information at several levels, in which each layer corresponds to a particular category of relevant information. The addition of a linguistic annotation layer allows also developing special ways of encoding, processing, and searching for linguistic variation phenomena, as the PapyGreek (formerly Sematia) platform developed at the University of Helsinki significantly demonstrates.[103] Each

101 Horsley 1994, 53–4.

102 To some extent, the Leiden+/XML markup of Papyri.info is a kind of non-linguistic, rather semantic annotation.

103 See Vierros 2018 and the chapters authored by Sonja Dahlgren, Erik Henriksson, and Marja Vierros in this volume.

layer added to the basic edition of papyrus texts is a better representation of their complex cognitive network as communication media.

When turning to the literary and paraliterary papyri, philological issues merge with the linguistic issues discussed above, complicating the picture even more. Consider, for instance, literary variants due to Koine forms attested in the papyrus tradition only: e.g., P.Aberd. 124 = GMP I 1 (TM 63334), i 14 π]ήχεωc instead of the 'regular' Ionic form πήχεοc in Hippocrates' *De fracturis* 37,[104] or Φωκείων (P.Tebt. VII 1160 [TM 957074], 2; P.Lond.Lit. 6 [TM 60260], ix 29) instead of Φωκήων in Homer's *Iliad* II 525. These are clearly later variations of the original text due to its transmission in a different linguistic context, but how much are they mistakes? For the ancient scribes who transcribed them they certainly sounded correct – perhaps even more correct than the 'original' form. How to 'regularize' such occurrences? Is it even possible to speak of 'regularization' at all?

The Homeric case is instructive. Following the standard papyrological paper editions, Alessia Bovo in P.Tebt. VII 1160, 2 prints Φ]ῳκείων in the main text and "i.e. Φωκήων" in the apparatus; quite curiously, but significantly, H. J. M. Milne in P.Lond.Lit. 6, ix 29 prints Φωκε[ίων with no indication in the apparatus (Fig. 18). It is curious, because Milne usually records the *lectio* of the Homeric codices and other textual variations in the apparatus, yet it is significant of a Koine linguistic form that is perceived as 'regular' in the Greek of the papyri (needless to say, the variant is 'relegated' in the apparatus in the current critical editions of the Homeric poem). Linguistic variation always bears broader cultural significance and 'substandard' forms very often betray cultural interferences that deserve more care than correcting 'irregular' forms. And what could we say of two further papyrological instances of the same verse, where the word under discussion is unfortunately lost in a material gap (BKT V.1 p. 4 [TM 60630], 8; P.Mich. inv. 6239 [TM 60527], 30)[105] and was supplied by the editors with the 'regular' Homeric form?

With such considerations in mind, let us turn to pure philological variants. Not rarely at all do the papyri preserve different readings from the main manuscript tradition, and often from the manuscript tradition *tout court*. Verse 125, again from the second book of the *Iliad*, is traditionally reported as Τρῶας μὲν λέξασθαι ἐφέστιοι ὅσσοι ἔασιν, which is what the manuscript tradition consistently transmits. P.Tebt. I 4 = VII 1159 (TM 61195, Tebtunis, 2nd cent. BC), which preserves lacunose portions of *Il.* II 95–211, has recently revealed a different reading for that verse: Τρῶες. Such a variant was mentioned by Eustathius of Thessalonica (12th cent.), in his famous commentary on the

104 There is much debate about the Greek employed in the treatises belonging to the Hippocratic corpus, since Hippocratic Ionic dialect is a literary language with several peculiarities (see Hanson 1970, 218 n. 25; Jouanna 2002, 133–55), but the genitive ending -εοc is genuine: see Jouanna 2002, 142.

105 The Berlin papyrus has been published by Müller 1995, 2–3; The Michigan fragment, which joins P.Aberd. 145, is published by Schwendner 1988, no. 2. In the latter, a second instance of Φωκήων (Hom. *Il.* II 533) is supplied in a lacuna (l. 38) as well.

Iliad, as found "in some copies" of Homer,[106] perhaps derived from Aristarchus' edition, which is roughly contemporary with the cited papyrus. The reading in the nominative was preferred by some modern editors to the accusative for such stylistic reasons as the parallel with the nominative ἡμεῖc in v. 126.[107] The papyrus is thus the oldest direct witness of this potentially correct reading, along with an unpublished piece from Oxyrhynchus, cited by West *ad l.*, which is reported to show the *epsilon* erased, perhaps in an attempt to correct the variant. In cases – far from infrequent – like this, where the oscillation between two equally valid forms disorients both modern and ancient scholars, how is it possible to restore an original text with absolute correctness? The papyrological evidence gives material substance to ancient textual circulation and poses important theoretical questions about textual transmission. If the text we seek to recover was not always the text actually circulating in ancient times, then we should at least distinguish two different critical attitudes: the philological path toward the reconstruction of the archetype beyond text transmission and the phenomenological path heading to a critical representation of the single steps of text transmission.

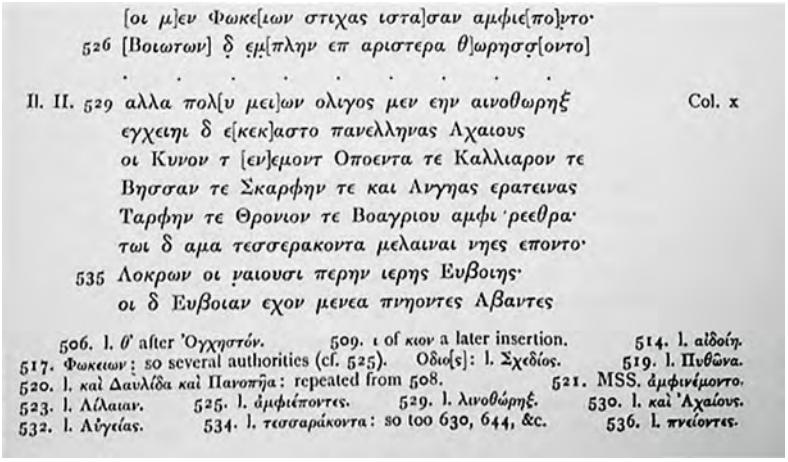

[οι μ]εν Φωκε[ιων στιχας ιστα]σαν αμφιε[πο]ντο·
[Βοιωτων] δ̣ εμ̣[πλην επ αριστερα θ]ωρησσ̣[οντο]
. . . . . .
Il. II. 529 αλλα πολ[υ μει]ων ολιγος μεν εην αινοθωρηξ Col. x
εγχειηι δ ε[κεκ]αστο πανελληνας Αχαιους
οι Κυνον τ [ευ]εμοντ Οποεντα τε Καλλιαρον τε
Βησσαν τε Σκαρφην τε και Ανγηας ερατεινας
Ταρφην τε Θρονιον τε Βοαγριου αμφι ῥεεθρα·
τωι δ αμα τεσσερακοντα μελαιναι νηες εποντο·
Λοκρων οι γαιουσι περην ιερης Ευβοιης·
οι δ Ευβοιαν εχον μενεα πνεοντες Αβαντες
506. l. ὅ' after 'Ογχηστόν. 509. ι of κιον a later insertion. 514. l. αἰδοίη.
517. Φωκειων: so several authorities (cf. 525). Οδιο[ς]: l. Σχεδίος. 519. l. Πυθῶνα.
520. l. καὶ Δαυλίδα καὶ Πανοπῆα: repeated from 508. 521. MSS. ἀμφινέμοντο.
523. l. Λίλαιαν. 525. l. ἀμφιέποντες. 529. l. λινοθώρηξ. 530. l. καὶ Ἀχαίους.
532. l. Αὐγείας. 534. l. τεσσαράκοντα: so too 630, 644, &c. 536. l. πνείοντες.

Fig. 18: Printed edition of P.Lond.Lit. 6.

106 Eust. *ad Hom. Il.* II 125–126, 190, 18–22 (I 291, 20–25 Valk) σημείωσαι δὲ καὶ τὴν καινότητα τοῦ σχήματος τοῦ «Τρῶας μὲν λέξασθαι, ἡμεῖς δ' ἐς δεκάδας διακοσμηθεῖμεν». ὤφειλε μὲν γὰρ εἰπεῖν· Τρῶας μὲν λέξασθαι, ἡμᾶς δὲ διακοσμηθῆναι. ὁ δὲ καινότερον ἐσχημάτισεν οὐκ ἀκολούθως τῷ Τρῶας ἐπενεγκὼν τὸ ἡμεῖς. διὸ ἔν τισιν ἀντιγράφοις εὕρηται, φασί, «Τρῶες μέν», ἵνα ὁμοίως ἐπάγηται τὸ ἡμεῖς δέ. See philological discussion in Isabella Bonati's commentary to P.Tebt. VII 1159 (forthcoming).

107 Leaf 1900; Ludwich 1902; West 1998. Conversely, Monro – Allen 1920, Allen 1931, Van Thiel 2010 print the other way around.

There are even more complex instances. P.Oslo inv. 1576 (TM 68976), a fragment of a catechism dealing with tumour-like diseases, partly overlaps with the text of P.Oxy. LXXX 5239 (TM 388545; both 2nd/3rd cent. AD).[108] As far as the extant parallel text is concerned, the wordings diverge from each other only for one variant: ὑδροκήλη (P.Oslo, l. 5) vs [ὑ]γροκήλη (P.Oxy., l. 15). The latter is usually considered as a minority variant (LSJ, quoting Poll. IV 203) of the former, but it is in fact used three times in medical literature.[109] Are we facing a trivialization in the Oslo papyrus, a simple phonetic variant in the Oxyrhynchus papyrus, or just two different traditions bearing the same degree of 'correctness', attesting to a fluid notion of technical language?

The embarrassment in handling such textual fluctuations, in which it is not actually possible to identify a 'correct' or even a 'standard' form as opposed to a 'deviant' variation, but just single instances in the text transmission, is just increased by those papyri where the ancient scribe himself added a divergent text. When the appended text is not found in the manuscript tradition the issue is even more puzzling: did the scribe want to correct a corrupted circulating text or just to record an equally valid circulating version of its? Examples of that are not rare among the papyri. For instance P.Tebt. II 272v (TM 60048), a late-2nd cent. AD fragment of Herodotus Medicus' *De remediis* describing the symptomatology of thirst and its treatment, the text of which corresponds in part to an excerpt preserved with Oribasius' treatment of thirst in case of fever (*Coll.med.* V 30, 6–7 = CMG VI 1, 1 Raeder). At l. 5, where the text reads αἰτίαι τῆc προσφορᾶc introducing the different reasons for giving the sick something to drink, the scribe adds two groups of three letters between dots above the line: *τῶν* above τῆc, and *ρῶν* above ρᾶc. Since nothing appears deleted, it is not clear if the scribe wanted to correct the text or just juxtapose two different versions of the same passage. We would have a scribe correcting the form unanimously preserved by the manuscript tradition and replacing it with an unattested variant. The P.Tebt. editors speak of "correction or alternative reading," Marie-Hélène Marganne of "hésitation."[110] If we should define it, we ought to call it a 'scribal variant'. And once we define it like that, one main question arises: which is the 'archetype', and which the 'variant'? And how to encode this situation in a digital edition?

Some test cases discussed by Federico Boschetti,[111] in the attempt to provide the literary databases such as the TLG with the due critical apparatuses, might direct our steps towards the enhancement of the extant markup strategies with tags specifically oriented to treat textual fluctuations, either philological or linguistic. This would mean to connect more information to the text encoded in the database, i.e., to add annotation layers. In Boschetti's examples, each word is identified by a unique number; the vari-

108 See Maravela – Leith 2007.

109 Orib. *Syn.Eust.* III 28, 6 and 9 = CMG VI 3, p. 75, 15–16 and 21 Raeder; Steph. *In Hp. Progn.* II 1 = CMG XI 1, 2, p. 140, 25 Duffy.

110 Marganne 1981, 76.

111 Boschetti 2007.

ants are aligned (i.e., linked) word by word; further information such as the origin of the alternative reading is provided within its own tag.[112]

This perspective would bring innovative solutions to the current question of how to manage the annotation of variants (either linguistic or philological) in the papyri. If we cease to consider a variant as a 'deviant version' to be 'corrected' and 'regularized', we can overcome the deadlock by looking at the full set of variants as a network or a dynamic system of textual transmission, and by thinking the digital edition as a multilayer text (which is somehow different from a multitext), a place for a dynamic collation of several versions stratified in the time or even at the same time level, so to valorise the materiality of the message in its transmission context.

As Monica Berti puts it,

> collecting multiple critical editions of the same text means building [...] a 'network of versions with a single, reconstructed root', so that scholars can compare different textual choices and conjectures produced by philologists. This process involves a new way of conceiving literary criticism because it produces a representation and visualization of textual transmission completely different from print conventions, where the text that is reconstructed by the editor is separated from the critical apparatus that is printed at the bottom of the page. In addition, the inclusion of images of manuscripts, papyri, and other source materials allows the reader to have a dynamic visualization of the textual tradition and to perceive the different channels of both the transmission and philological production of the text that is usually hidden in the static, concise, and necessarily selective critical apparatuses of standard printed editions. Producing a multitext, therefore, means producing multiple versions of the same text, which are the representation of the different steps of its transmission and reconstruction, from manuscript variants to philological conjectures. This process has fundamental consequences for the study of ancient sources in general and for fragmentary ones in particular, given that, while studying fragments and evaluating their distance from the original version, it is imperative to examine the manuscript variants of the source text, in order to see what can be attributed to the witness or to the transmission of the text across centuries.[113]

It is a completely new and different way of considering the criticism of ancient texts. The texts – in our case, the papyri – become meta-texts (meta-papyri) and the critical apparatus dissolves in a network of references, connections, and versions. This kind of textual network has also much to do with the ancient philological care as testified by numerous instances: as it is shown by the abovementioned 'scribal variants' and as it will be resumed in section 5, ancient 'philology' was much more interested in a fluid textual transmission (the "accretive model of composition" claimed by Ann Hanson and mentioned in section 3) rather than in the fixation of a stable (and static), canonical (and constrained), 'correct' version of the texts.

112 This is not without issues, such as the problem of how to tag broken words.

113 Berti 2010, 4–5.

5 The ancient document as a hypertext

Non ci si può bagnare
due volte nello stesso fiume,
né prevedere i cambiamenti di costume.
E intanto passa ignaro
il vero senso della vita.

Franco Battiato, *Di passaggio*.[114]

The concept of 'liquid modernity',[115] as first introduced by sociologist Zygmunt Bauman, is the characterization of today's highly developed global societies that produce increasing feelings of uncertainty in the individuals, and fluid networks instead of rigid categorizations. This is by no means a contemporary concept, since – as is of common knowledge – already presocratic philosopher Heraclitus coined the famous phrase πάντα ῥεῖ to describe the incessant flow of universal existence: "one cannot wet himself in the same river twice," according to his equally famous metaphor.

At any rate, the digital shift contributed to increase the sensation of modern liquidity in textual matters too. What Primo Levi wittily described in 1984 in his short essay *Lo scriba* ("The scribe" – meaningful title, by the way) is just enhanced by the most recent developments in digital textuality:

> Due mesi fa, nel settembre 1984, mi sono comprato un elaboratore di testi, cioè uno strumento per scrivere che va a capo automaticamente a fine riga, e permette di inserire, cancellare, cambiare istantaneamente parole o intere frasi; consente insomma di arrivare d'un colpo ad un documento finito, pulito, privo di inserti e di correzioni. [...] Ho notato che scrivendo così si tende alla prolissità. La fatica di un tempo, quando si scalpellava la pietra, conduceva allo stile "lapidario": qui avviene l'opposto, la manualità è quasi nulla, e se non ci si controlla si va verso lo spreco di parole. [...] Qui tu scrivi, le parole appaiono sullo schermo nitide, bene allineate, ma sono ombre: sono immateriali, prive del supporto rassicurante della carta. "La carta canta", lo schermo no; quando il testo ti soddisfa, lo "mandi su disco", dove diventa invisibile. C'è ancora, latitante in qualche angolino del disco-memoria, o l'hai distrutto con qualche manovra sbagliata? [...] Un amico letterato mi obietta che così va perduta la nobile gioia del filologo intento a ricostruire, attraverso le successive cancellature e correzioni, l'itinerario che conduce alla perfezione dell'Infinito: ha ragione, ma non si può aver tutto.[116]

114 ["One cannot bathe / twice in the same river, / nor foresee changes in customs. / And meanwhile, the true meaning of life / passes by unnoticed."]

115 This section develops a hitherto unpublished talk, titled "Il documento antico come ipertesto", delivered within the doctoral seminar "DH – Digital Humanities" at the University of Parma on October 23, 2018. I warmly thank Massimo Magnani for inviting me to participate in that seminar.

116 Levi 2017, 841–4. ["Two months ago, in September 1984, I bought a word processor, that is, a writing tool that automatically moves to the next line at the end of the current one and allows you to insert, delete, instantly change words or entire sentences; in short, it enables you to produce a finished, clean document free of inserts and corrections in one go. [...] I have noticed that writing in this way tends to lead to

We perceive the digital volatility of text as subverting because in our collective unconscious we have an idea of 'written text' as fixed and stable, unchanging and unchanged, and of 'book' as a container of a canonical and immutable aspect of writing. In fact, the creation of 'canons' is an ideological/religious need (from the most known ancient examples: the Peisistratid canon of Homer served political ideology; the Hippocratic corpus fit claims of authority; the Bible fixed a religious order, first Hebraic and then Christian), not a practical or cultural one. Canons support authority. We have several ancient cases of the struggle between writing as support to oral communication and writing as authorial fixation of a canon, or of a text:[117] Plato vs the Sophists, or the transition between Herodotus' public lectures to Thucydides' κτῆμα ἐς αἰεί. The most relevant criticism claimed by the opponents to 'fixed' writing was that a written text is not interactive – so, for instance, in Plato's *Phaedrus* and Alcidamas' *On the Sophists*:

> Writing, Phaedrus, has this strange quality, and is very like painting; for the creatures of painting stand like living beings, but if one asks them a question, they preserve a solemn silence. And so it is with written words; you might think they spoke as if they had intelligence, but if you question them, wishing to know about their sayings, they always say only one and the same thing. And every word, when once it is written, is bandied about, alike among those who understand and those who have no interest in it, and it knows not to whom to speak or not to speak; when ill-treated or unjustly reviled it always needs its father to help it; for it has no power to protect or help itself.[118] (Plat. *Phaedr*. 275d–e)

> Written discourses, in my opinion, certainly ought not to be called real speeches, but they are as wraiths, semblances, and imitations. It would be reasonable for us to think of them as we do of bronze statues, and images of stone, and pictures of living beings; just as these last mentioned are but the semblances of corporeal bodies, giving pleasure to the eye alone, and are of no practical value, so, in the same way, the written speech, which employs one hard and fast form and arrangement, if privately read, makes an impression, but in crises, because of its rigidity, confers no aid on its possessor. And, just as the living human body has far less comeliness than a beautiful statue, yet manifold practical service, so also the speech which comes directly from the mind, on the spur of the moment, is full of life and action, and keeps pace with the events like a real person, while the written discourse, a mere semblance of the living speech, is devoid of all efficacy.[119] (Alcid. *Soph*. 27–28)

---

verbosity. The labour of the past, when one chiselled into stone, led to a 'lapidary' style: here, the opposite happens, manual effort is almost non-existent, and if one is not careful, it leads to a waste of words. […] Here you write, the words appear on the screen clearly, well-aligned, but they are shadows: they are immaterial, lacking the reassuring support of paper. 'Paper sings,' the screen does not; when the text satisfies you, you 'send it to disk,' where it becomes invisible. Is it still there, lurking in some corner of the memory disk, or have you destroyed it with some wrong command? […] A writer friend objects that this way, the noble joy of the philologist intent on reconstructing, through successive deletions and corrections, the path that leads to the perfection of the Infinite is lost: he is right, but you cannot have everything."]

117 See Reggiani 2023b.

118 Transl. H. N. Fowler, from the Perseus Digital Library.

119 Transl. L. Van Hook, from Attalus (https://www.attalus.org/translate/alcidamas.html).

In fact, we have several possible examples of ancient texts that are effectively interactive in their everyday materiality. They do interact with their readers: by means of commentaries, intertextual cross-references, critical and diacritical markup or layout elements that enhance the conveyed message as a sort of written counterpart of oral gesturality; most interestingly, blank spaces offered the possibility to add answers or personal/official annotations and to increase, modify, update the text.[120] They interact with their authors as well (annotations, comments, collated variants, corrections…) and even with the material context (linguistic variation and change, material support, fragmentation). All the topics discussed in the previous sections 2-3-4 are in fact instances of interactions.

Galen – whose thorough activity as a 'philologist' is widely known thanks to his own personal testimony[121] – did complain about text fluctuations, but in his compiling work he did not report a canon, he did rather collate the copies and report the main variants, as well as the textual additions he might have found.[122] Not much dissimilar could have been the editorial practice in the Library of Alexandria itself. According to some scholars, the Alexandrian philologists used to choose carefully a basic copy of a text – a circulating copy – and to work on that, adding critical marks, appending marginal notes and variants, deleting spurious passages, and eventually discussing interpretations in the commentary (*hypomnema*).[123] It was a kind of open text, a hypertext in a sense, because it interacted with its creators and its users in various ways at several different levels of textuality, and – from another viewpoint – it was a 'liquid' text, subject to transformation over the time.

It is sufficiently clear that an open edition is needed in order to critically represent an open text, otherwise we will unavoidably lose relevant information. The status of papyrology as a 'liquid philology', as presented in section 1, can help understanding and

120 See Luiselli 2008, 708; Sarri 2018, 14, 145.

121 See Totelin 2009; Roselli 2012 and 2020, with earlier bibliography.

122 E.g.: Gal. *Comp.med.loc.* I 2 (XII 400, 7–11 Kühn) "Heras literally wrote as follows: 'one part of rocket seeds, one part of cardamom, one part of sodium carbonate. In some copies, it is written simply 'one part of rocket, one part of cardamom, one part of sodium carbonate'"; *Comm.Hp.Off.* III 22 (XVIIIb 863, 15–864, 5 Kühn) "In such books, which contain an interpretation of many things in an abbreviated manner, it often happens that the author writes about the same things in different ways, intending to use certain words more than others. Then, the copyist finds some of these written in the margins and some written on the front of the book, and writes them all at the end of the manuscript, where they will appear to be arranged in the most reasonably best way possible"; *Comm.Hp.Epid.* IV 21 (XVIIb 194, 11–195, 3 Kühn) "But I did not find this recipe in some of the other copies, nor does any of the commentators know of it, except that Dioscorides wrote it on the front of the book, having found it in only two copies, which somehow had the same formulation: "For a warm nature, in a warm season, sleeping in the cold makes you gain weight, sleeping in the heat makes you lose weight." In fact, we ourselves have found this formulation in all the copies we have read, having intentionally examined everything in public libraries and those of our friends" (transl. mine).

123 See Irigoin 1994, 42, with the comment by Montanari 1994, 85.

answering this claim. As in Primo Levi's reflection, there is a diffused feeling that the hypertext is challenging the *Urtext* model,[124] though responses differ from each other. While multitext is a "method to track multiple versions of a text across time",[125] it is possible to envisage "a more holistic notion", where the user can access "not only [...] a presentational publication layer but also by allowing access to the underlying encoding of the repository or database beneath", a "critical edition, with sources fully incorporated, [which] would potentially provide an interactive resource that assists the user in creating virtual research environments", and which would relieve an editor from making "any authoritative decisions that supersede all alternative readings if all possibilities can be unambiguously reconstructed from the base manuscript data".[126] As I contended elsewhere, with a practical example referred to the *Michigan Medical Codex*,[127] the model that better describes this ideal condition is perhaps an ontology, since "the use of stand-off metadata encoded within ontology allows us to express an open-ended number of interpretations, whereas a markup-based solution would not make this possible due to obvious reasons of overlapping hierarchies."[128]

The digital papyrus is a different entity than the 'traditional' papyrus. It has its own ontology that can produce a completely different textual criticism, thanks to the new virtual medium where it is represented. The need to reconstruct and print some 'canonical' text, which is ultimately connected to a paper-like way of thinking, simply dissolves in the multi-dimensional, meta-dimensional, and tabular digital space. The digital document is no more a product of philological interpretation, but a new, enhanced avatar of the original document and of all its metatextual and intertextual connections and networks – all its dispositive, in foucaldian terms, or also what has recently been referred to the notion of 'multimodality'.[129] It is a meta-papyrus in a new virtual materiality, fruit of a digital interpretation, and the digital critical edition positions itself, beyond the apparatus, as a further step in text transmission (see Figs. 19–20).

Where do we find, then, criticism in all of this? Of course, we do not have to think that digital editions should be uncritical or agnostic. We must recall that "encoding a text is an interpretive act"[130] by itself: and this is even truer if we consider that the

124 The expression is borrowed from Bolter 1991. It is worth recalling the interesting observation by Cayless 2010, 162, that traditional commentary is a hypertext in print (see also *ibid.*, 170). See also Clivaz 2012, who – besides her own observations – cites Umberto Eco's opinion that the digital age may mean the end of the history of variants and of the notion of 'original text' (Eco – Origgi 2003).

125 Babeu 2011, 214.

126 Bodard – Garcés 2009, 96.

127 For this papyrus see above, §3.

128 Romanello – Berti – Boschetti – Babeu – Crane 2009, 158. See Reggiani 2017, 266–8; 2018a, 5–6 and 48–53; 2019a, 353–5; 2019d. An ontology is a way of showing the properties of a subject and how they are related, by defining a set of terms and relational expressions that represent the entities in that subject.

129 See the chapter authored by Klaas Bentein in this volume.

130 Owens 2011; cf. Tarte 2011c, 1. On the critical outcome of computational tools see also the notion of "algorithmic criticism" as outlined by Ramsay 2011.

encoder (the digital papyrologist) must employ as much criticism and careful discernment as possible in order to give the papyrological object its correct digital representation. Encoding means adapting the printed conventions to the new digital medium, following strict computational standards. Digital criticism seems to mean interpreting both the papyrological data (the object, its text, its context) and the printed critical edition(s) in order to produce a digital representation of the papyrus as a metatextual and multimodal dispositive, i.e., an interconnected and multidimensional network of text, intertexts and other transtextual layers, metadata, image, and so on.

Needless to say, although computers indeed challenge the idea of the 'authority' of the editor, they do create at the same time a new much more complex form of 'authority':

> It is clear that these media, when used within a wider intellectual perspective as a cognitive tool for research and instruction and not only as a pragmatic medium that can 'do certain things for us', can challenge and redefine our notions of 'text', textuality, and text transmission.[131]

Concept that digital papyrology redefines the notion of 'text' is embedded in the consideration that electronic technologies offer a completely new room to scholarly research. The digital space does totally change scholarly parameters. We do not deal with texts any more: we deal with meta-texts, hypertexts, multi-layer annotated texts enriched by metadata, embedded apparatuses – virtual entities that are subject to quick – which does not mean arbitrary – updates, to a constant renovation, to a continuous scholarly labour. Thence, an unavoidable fact: "We need to move in the direction of digitally conceived and initiated types of information and away from mopping up information from print sources."[132]

The following statement by Greg Crane must unavoidably be kept in mind: "in a digital age, philologists need to treat our editions as components of larger, well-defined corpora rather than as the raw material for printed page layouts".[133] And since hypertextual multi-layer architectures seem to be the best way to respond to these needs, and they look so close to the interactive dynamics of ancient texts, which they represent at best, we can safely conclude that a proper digital critical edition can become (also) a further step in textual transmission.[134]

---

131 Gagos 2001, 515 n. 8.

132 Bagnall – Gagos 2007, 74. See also Purpura 2001, 5: "i problemi sono oggi connessi alla difficoltà di abbandonare rapidamente radicati atteggiamenti connessi all'opera cartacea" ["The problems today are related to the difficulty of quickly abandoning deeply rooted attitudes associated with paper-based work"]; Romanello – Berti – Boschetti – Babeu – Crane 2009, 165: ("Once we are able to overcome the physical limits of printed editions by joining together variants and conjectures referring to the same texts, it also becomes possible to look at the texts from a new and broader perspective, with possible consequences for our knowledge and comprehension of them"; Cayless 2010, 148: "perhaps emphasis on technology that faithfully replicates the printed appearance of documents is misplaced".

133 Crane 2010.

134 See Reggiani 2022a.

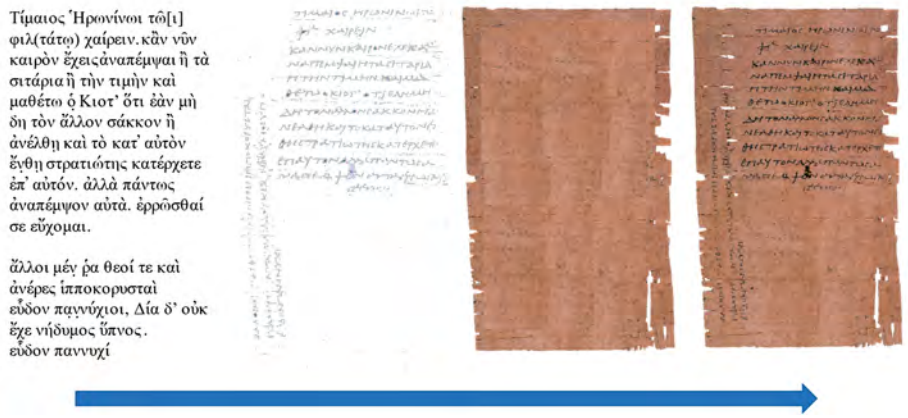


Fig. 19: The main components of the materiality of P.Flor. II 259 (see above): the text (here already in its editorial representation), its palaeographical appearance, the material support, resulting in the papyrus as a communicative artefact.

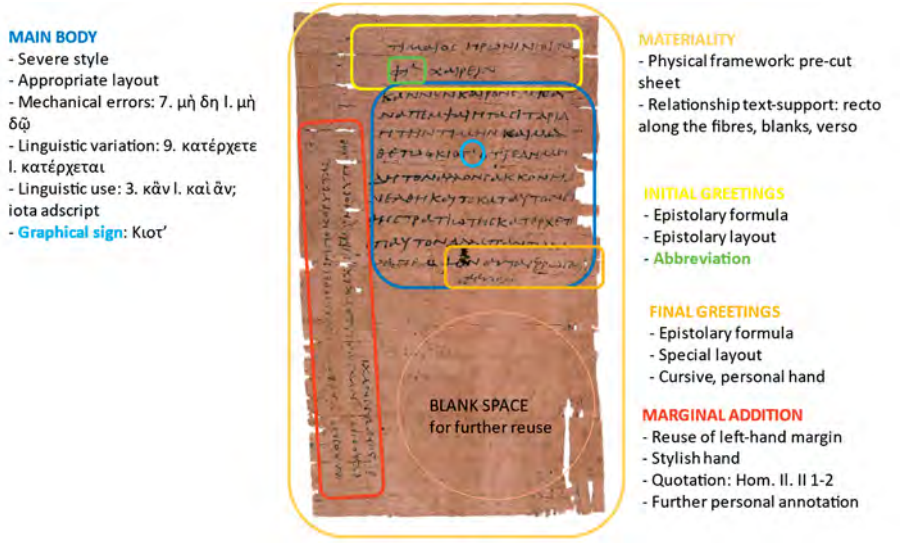


Fig. 20: Various possible textual, transtextual and material layers of P.Flor. II 259 (see above).

Bibliography

Allen, T. W. (1931), *Homeri Ilias, II Libros I-XII continens*, Oxford.

Andorlini, I. (1993), *L'apporto dei papiri alla conoscenza della scienza medica antica*, in *Aufstieg und Niedergang der Römischen Welt*, ed. by W. Haase – H. Temporini, Berlin – New York, II 37.1, 458-562.

Andorlini, I. (2017), *πολλὰ ἰατρῶν ἐστι συγγράμματα. Scritti sui papiri e la medicina antica*, ed. by N. Reggiani, Florence.

Ast, R. – Choat, M. – Cromwell, J. – Lougovaya, J. – Yuen-Collingridge, R. (2021), eds., *Observing the Scribe at Work. Scribal Practice in the Ancient World*, Leiden.

Ast, R. – Essler, H. (2018), Anagnosis*, Herculaneum and the* Digital Corpus of Literary Papyri, in Reggiani 2018b, 63–73.

Babeu, A. (2011), *"Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classics*, Washington DC.

Baetens, G. (2020), *A Survey of Petitions and Related Documents from Ptolemaic Egypt*, Leuven, https://www.trismegistos.org/dl.php?id=18.

Bagnall, R. S. – Gagos, T. (2007), *The Advanced Papyrological Information System: Past, Present, and Future*, in *Proceedings of the 24th International Congress of Papyrology (Helsinki, 1–7 August 2004)*, ed. by J. Frosen – T. Purola – E. Salmenkivi, Helsinki, I, 59–74

Bauman, Z. (2000), *Liquid Modernity*, Malden.

Bauman, Z. (2007), *Liquid Times. Living in an Age of Uncertainty*, Malden.

Bauman, Z. (2011), *Culture in a Liquid Modern World*, Malden.

Berkes, L. (2018), *Perspectives and Challenges in Editing Documentary Papyri Online. A Report on Born-Digital Editions through Papyri.info*, in Reggiani 2018b, 75–85.

Berti, M. (2010), *Fragmentary Texts and Digital Libraries*, http://www.fragmentarytexts.org/wp-content/uploads/2010/07/Berti-Fragmentary-Texts-and-Digiital-Libraries.pdf.

Bodard, G. – Garcés, J. (2009), *Open Source Critical Editions: A Rationale*, in *Text Editing, Print, and the Digital World*, ed. by M. Deegan – K. Sutherland, Farnham – Burlington, 84–98.

Bolter, J. D. (1991), *The Computer, Hypertext, and Classical Studies*, American Journal of Philology 112, 541–5.

Bongiovanni, R. (2024), *Ricettari medici e rimedi magici: un rapporto in evoluzione*, in Reggiani 2024a, 169–88.

Boschetti, F. (2007), *Methods to Extend Greek and Latin Corpora with Variants and Conjectures: Mapping Critical Apparatuses onto Reference Text*, in: *Proceedings of the Corpus Linguistics Conference CL 2007 (Birmingham 2007)*, article #150, http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/150Paper.pdf.

Bucking, S. (2007), *On the Training of Documentary Scribes in Roman, Byzantine, and Early Islamic Egypt: A Contextualized Assessment of the Greek Evidence*, ZPE 159, 229–47.

Canfora, L. (2002), *Il copista come autore*, Palermo.

Cayless, H. A. (2010), *Ktêma es aiei: Digital Permanence from an Ancient Perspective*, in *Digital Research in the Study of Classical Antiquity*, ed. by G. Bodard – S. Mahony, Farnham – Burlington, 139–50.

Choat, M. (2006), *Echo and Quotation of the New Testament in Papyrus Letters to the End of the Fourth Century*, in *New Testament Manuscripts. Their Texts and Their World*, ed. by T. J. Kraus – T. Nicklas, Leiden – Boston, 267–92.

Choat, M. (2021), *Text and Paratext in Documentary Papyri from Roman Egypt*, in Ast – Choat – Cromwell *et al.* 2021, 281–98.

Clivaz, C. (2012), *Homer and the New Testament as "Multitexts" in the Digital Age*, Scholarly and Research Communication 3, https://doi.org/10.22230/src.2012v3n3a97.

Coles, R. A. (1966), *Reports of Proceedings in Papyri*, Bruxelles.

Collins, D. (2008), *The Magic of Homeric Verses*, Classical Philology 103, 211–36.

Crane, G. (2010), *Give Us Editors! Re-inventing the Edition and Re-thinking the Humanities*, in *Online Humanities Scholarship: The Shape of Things to Come. Proceedings of the Mellon Foundation Online Humanities Conference at the University of Virginia (March 26–28)*, ed. by J. McGann, Houston, 137–70.

Cribiore, R. (1996), *Writing, Teachers, and Students in Graeco-Roman Egypt*, Atlanta.

Cribiore, R. (2001), *Gymnastics of the Mind: Greek Education in Hellenistic and Roman Egypt*, Princeton – Oxford.

Cribiore, R. (2019), *Genetic Criticism and the Papyri: Some Suggestions*, in *Greek Medical Papyri. Text, Context, Hypertex*, ed. by N. Reggiani, Berlin – Boston, 173–91.

de Haro Sanchez, M. (2015), *Between Magic and Medicine: The Iatromagical Formularies and Medical Receptaries on Papyri Compared*, ZPE 195, 179–89.

Del Corso, L. (2018), *I rotoli dell'*Athenaion Politeia *nel contesto della produzione libraria dell'Egitto greco-romano*, in *Athenaion Politeiai tra storia, politica e sociologia: Aristotele e Pseudo-Senofonte*, ed. by C. Bearzot – M. Canevaro – T. Gargiulo – E. Poddighe, Milan, 33–55.

Depauw, M. – Stolk, J. (2015), *Linguistic Variation in Greek Papyri: Towards a New Tool for Quantitative Studies*, GRBS 55, 196–220.

Dover, K. J. (1997), *The Evolution of Greek Prose Style*, Oxford.

Eckerman, C. (2010), *Hexameters from Late Antiquity with a Homeric Allusion*, BASP 47, 29–32.

Eco, U. – Origgi, G. (2003), *Auteurs et autorité : Un entretien avec Umberto Eco*, in *Texte-e: Le texte à l'heure de l'Internet*, ed. by G. Origgi – N. Arikha, Paris, 215–30.

Evans, T. V. (2007), *Greetings from Alexandria*, in *Proceedings of the 24th International Congress of Papyrology (Helsinki, 1–7 August, 2004)*, ed. by J. Frösén – T. Purola – E. Salmenkivi, Helsinki, I, 299–308.

Evans, T. V. (2010a), *Identifying the Language of the Individual in the Zenon Archive*, in *The Language of the Papyri*, ed. by T. V. Evans – D. D. Obbink, Oxford, 51–70.

Evans, T V. (2010b), *Standard Koine Greek in Third Century BC Papyri*, in *Proceedings of the Twenty-Fifth International Congress of Papyrology (Ann Arbor, July 29–August 4, 2007)*, ed. by T. Gagos, Ann Arbor, 197–206.

Evans, T. V. (2012a), *Complaints of the Natives in a Greek Dress. The Zenon Archive and the Problem of Egyptian Interference*, in *Multilingualism in the Graeco-Roman Worlds*, ed. by A. Mullen – P. James, Cambridge, 106–23.

Evans, T. V. (2012b), *Linguistic and Stylistic Variation in the Zenon Archive*, in *Variation and Change in Greek and Latin*, ed. by M. Leiwo – H. Halla-Aho – M. Vierros, Helsinki, 25–42.

Fleischer, K. (2021), *Die Papyri Herkulaneums im Digitalen Zeitalter: Neue Texte durch neue Techniken – eine Kurzeinführung*, Berlin – Boston.

Fournet, J.-L. (2004), *Entre document et littérature : la pétition dans l'Antiquité tardive*, in *La pétition à Byzance. XXe Congrès international des Études byzantines, 19-25 août 2001. Table ronde*, ed. by D. Feissel – J. Gascou, Paris, 61–74.

Fournet, J.-L. (2007), *Disposition et réalisation graphique des lettres et des pétitions protobyzantines : pour une paléographie « signifiante » des papyrus documentaires*, in *Proceedings of the 24th International Congress of Papyrology (Helsinki, 1–7 August, 2004)*, ed. by J. Frösén – T. Purola – E. Salmenkivi, Helsinki, I, 353–67.

Fournet, J.-L. (2012), *Homère dans les papyrus non littéraires : le Poète dans le contexte de ses lecteurs*, in *I papiri omerici. Atti del convegno internazionale di studi (Firenze, 9–10 giugno 2011)*, ed. by G. Bastianini – A. Casanova, Florence, 12–57.

Fournet, J.-L. (2018), *Archives and Libraries in Greco-Roman Egypt*, in *Manuscripts and Archives: Comparative Views on Record-Keeping*, ed. by A. Bausi – C. Brockmann – M. Friedrich – S. Kienitz, Berlin – Boston, 171–200.

Fournet, J.-L. (2022), *Beyond the Text or the Contribution of "Paléographie signifiante" in Documentary Papyrology: The Example of Formats in Late Antiquity*, in *Novel Perspectives on Communication Practices in Antiquity: Towards a Historical Social-Semiotic Approach*, ed. by K. Bentein – Y. Amory, Leiden – Boston, 17–28.

Gagos, T. (2001), *The University of Michigan Papyrus Collection: Current Trends and Future Perspectives*, in *Atti del XXII Congresso Internazionale di Papirologia (Firenze, 23–29 agosto 1998)*, ed. by I. Andorlini – G. Bastianini – M. Manfredi – G. Menci, Florence, II, 511–37.

Gazza, V. (1955), *Prescrizioni mediche nei papiri dell'Egitto greco-romano, I*, Aegyptus 35, 86–110.

Genette, G. (1992), *The Architext: An Introduction*, Berkeley. [Or. ed. *Introduction à l'architexte*. Paris 1979]

Genette, G. (1997), *Palimpsests. Literature in the Second Degree*, Lincoln. [Or. ed. *Palimpsestes : la littérature au second dégrée*, Paris 1982]

Hanson, A. E. (1970), P. Antinoopolis *184: Hippocrates*, Diseases of Women, in *Proceedings of the XIIth International Congress of Papyrology (Ann Arbor, 13–17 August 1968)*, ed. by D. H. Samuel, Toronto, 213–22.

Hanson, A. E. (1997), *Fragmentation and the Greek Medical Writers*, in *Collecting Fragments / Fragmente Sammeln*, ed. by G. W. Most, Göttingen, 289–314.

Hanson, A. E. (2002), *Papyrology: A Discipline in Flux*, in *Disciplining Classics / Altertumwissenschaft als Beruf*, ed. by G.W. Most, Göttingen, 191–206.

Hanson, A. E. (2010), *Doctors' Literacy and Papyri of Medical Content*, in *Hippocrates and Medical Education*, ed. by M. Horstmanshoff, Leiden, 187–204.

Hoogendijk, F. A. J. – van Gompel, S. M. T. (2018), eds., *The Materiality of Texts from Ancient Egypt. New Approaches to the Study of Textual Material from the Early Pharaonic to the Late Antique Period*, Leiden – Boston.

Horsley, G. H. R. (1994), *Papyrology and the Greek Language. A Fragmentary Abecedarius of Desiderata for Future Study*, in *Proceedings of the 20th International Congress of Papyrologists (Copenhagen, 23–29 August, 1992)*, ed. by A. Bülow-Jacobsen, 48–70.

Irigoin, J. (1994), *Les éditions de textes*, in *La philologie grecque à l'époque hellénistique et romaine*, Vandoeuvres – Genève, 39–82.

Johnson, W. A. (2009), *Introduction*, in *Ancient Literacies: The Culture of Reading in Greece and Rome*, ed. by W. A. Johnson – H. N. Parker, Oxford, 3–10.

Jouanna, J. (2002), ed., *Hippocrate. La nature de l'homme* (Corpus Medicorum Graecorum I 1.3), 2nd ed., Berlin.

Laffi, U. (2013), *In greco per i Greci. Ricerche sul lessico greco del processo civile e criminale romano nelle attestazioni di fonti documentarie romane*, Pavia.

Lamé, M. (2014), *Primary Sources of Information, Digitization Processes and Dispositive Analysis*, in *Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem (Bologna, 18–19 September 2014)*, https://doi.org/10.1145/2802612.2802645.

Larsen, L. I. (2018), *School Texts*, in *A Companion to Late Antique Literature*, ed. by S. McGill – E. J. Watts, New York, 471–90.

Leaf, W. (1900), ed., *The Iliad, I. Books I-XII*, 2nd ed., London.

Levi, P. (2017), *Opere complete I-II*, ed. by M. Belpoliti, Turin.

Lord, A. B. (1960), *The Singer of Tales*, Cambridge MA.

Ludwich, A. (1902), ed., *Homeri Ilias,* I, Leipzig.

Luiselli, R. (1999), *A Study of High Level Greek in the Non-Literary Papyri from Roman and Byzantine Egypt*, PhD Diss., University College London.

Luiselli, R. (2008), *Greek Letters on Papyrus, First to Eighth Centuries: A Survey*, Asiatische Studien 62, 677–737.

Maas, P. (1960), *Textkritik*, 4. Auflage, Leipzig. [English translation of the 2nd German edition, by B. Flower: *Textual Criticism*, Oxford 1958]

Magnani, M. (2018), *The Other Side of the River. Digital Editions of Ancient Greek Texts Involving Papyrus Witnesses*, in Reggiani 2018b, 87–102.

Maravela, A. – Leith, D. (2007), *A Medical Catechism on Tumours from the Collection of the Oslo University Library*, in *Proceedings of the 24th International Congress of Papyrology (Helsinki 2004)*, ed. by J. Frösén – T. Purola – E. Salmenkivi, Helsinki, 637–50.

Marganne, M.-H. (1981), *Un fragment du medecin Herodote: P. Tebt. II 272*, in: *Proceedings of the Sixteenth International Congress of Papyrology (New York 1980)*, Chico, 73–8.

Mascellari, R. (2021), *La lingua delle petizioni nell'Egitto romano. Evoluzione di lessico, formule e procedure dal 30 a.C. al 300 d.C.*, Florence.

Meier, T. – Ott, M. R. – Sauer, R. (2015), eds., *Materiale Textkulturen: Konzepte, Materialien, Praktiken*, Berlin – Boston.

Messeri, G. (1998), *P.Flor. II 259*, in *Scrivere libri e documenti nel mondo antico*, ed. by G. Cavallo – E. Crisci – G. Messeri – R. Pintaudi, Florence, 208–9.

Migliardi Zingale, L. (2003), *Sull'uso dei formulari nella confezione di documenti giuridici: testimonianze dall'Egitto romano e bizantino*, in *Atti del VII Convegno Nazionale di Egittologia e Papirologia (Siracusa, 29 novembre – 2 dicembre 2001)*, ed. by C. Basile – A. Di Natale, Siracusa, 99–106.

Monro, D. B. – Allen, T. W. (1920), eds., *Homeri Opera*, *I Iliadis libros I-XII continens*, 3rd ed., Oxford.

Montanari, F. (1994), *Discussion* of Irigoin 1994, in *La philologie grecque à l'époque hellénistique et romaine*, Vandoeuvres – Genève, 83–7.

Monte, A. (2024), *Riflessioni sul formato dei papiri greci di contenuto farmacologico*, in Reggiani 2024a, 101–16.

Müller, W. (1995), *Ilias-Handschriften aus der Berliner Papyrus-Sammlung (IV)*, APF 41, 1–19.

Nachtergaele, D. (2013), *The Asklepiades and Athenodoros Archives: A Case Study of a Linguistic Approach to Papyrus Letters*, GRBS 53, 269–93.

Nachtergaele, D. (2016), *Variation in Private Letters: The Papyri of the Apollonios Strategos Archive*, GRBS 56, 140–63.

Nagy, G. (2010), *The Homer Multitext Project*, in *Online Humanities Scholarship: The Shape of Things to Come. Proceedings of the Mellon Foundation Online Humanities Conference at the University of Virginia (March 26-28, 2010)*, ed. by J. McGann – A. Stauffer – D. Wheeles – M. Pickard, Houston, 87–112. [https://chs.harvard.edu/curated-article/gregory-nagy-the-homer-multitext-project]

Owens, T. (2011), *Defining Data for Humanists: Text, Artifact, Information or Evidence?*, Journal of Digital Humanities 1.1, http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens.

Pasquali, G. (1988), *Storia della tradizione e critica del testo*, Florence. [1st ed. Florence 1934]

Piquette, K. E. (2018), *Revealing the Material World of Ancient Writing: Digital Techniques and Theoretical Considerations*, in Hoogendijk – van Gompel 2018, 94–118.

Porter, S. E. – Pitts, A. W. (2013), *The Disclosure Formula in the Epistolary Papyri and in the New Testament: Development, Form, Function, and Syntax*, in *The Language of the New Testament. Context, History, and Development*, ed. by S. E. Porter – A. W. Pitts, Leiden – Boston, 421–38.

Powell, B. B. (2014), ed., *Homer. The Iliad*, transl., introd., and notes by B. B. Powell, Oxford – New York.

Purpura, G. (2001), *Le nuove tecnologie informatiche applicate alla ricerca e allo studio del diritto romano e dei diritti dell'antichità*, Rivista di Diritto Italiano 1, 1–10.

Ramsay, S. (2011), *Reading Machines. Toward an Algorithmic Criticism*, Urbana – Chicago – Springfield.

Rathbone, D. (1991), *Economic Rationalism and Rural Society in Third-Century A.D. Egypt. The Heroninos Archive and the Appianus Estate*, Cambridge.

Reggiani, N. (2017), *Digital Papyrology I. Tools, Methods and Trends*, Berlin – Boston.

Reggiani, N. (2018a), *The* Corpus of the Greek Medical Papyri *and a New Concept of Digital Critical Edition*, in Reggiani 2018b, 3–61.

Reggiani, N. (2018b), ed., *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, Berlin – Boston.

Reggiani, N. (2018c), *I papiri greci di medicina come fonti storiche: il caso dei rapporti dei medici pubblici nell'Egitto romano e bizantino*, Aegyptus 98, 107–30.

Reggiani, N. (2019a), *La papirologia digitale. Prospettiva storico-critica e sviluppi metodologici*, Parma.

Reggiani, N. (2019b), *Linguistic and Philological Variants in the Papyri: A Reconsideration in the Light of the Digitization of the Greek Medical Papyri*, in *Greek Medical Papyri. Text, Context, Hypertex*, ed. by N. Reggiani, Berlin – Boston, 237-56.

Reggiani, N. (2019c), *The Corpus of Greek Medical Papyri Online and the Digital Edition of Ancient Documents*, in *Proceedings of the 28th Congress of Papyrology (Barcelona, August 1–6, 2016)*, ed. by A. Nodar – S. Torallas Tovar, Barcelona, 843–56.

Reggiani, N. (2019d), *Ancient Doctors' Literacies and the Digital Edition of Papyri of Medical Content*, Classics@ 17, https://classics-at.chs.harvard.edu/classics17-reggiani-2.

Reggiani, N. (2019e), *Transmission of Recipes and Receptaria in Greek Medical Writings on Papyrus. Between Ancient Text Production and Modern Digital Representation*, in *On the Track of the Books: Scribes, Libraries and Textual Transmission*, ed. by R. Berardi – N. Bruno – L. Fizzarotti, Berlin – Boston, 167–88.

Reggiani, N. (2019f), *Papirologia. La cultura scrittoria dell'Egitto greco-romano*, Parma.

Reggiani, N. (2020), *Digitizing Medical Papyri in Question-and-Answer Format*, in *Ancient Greek Medicine in Questions and Answers: Diagnostics, Dicactics, Dialectics*, ed. by M. Meeusen, Leiden – Boston, 181–212.

Reggiani, N. (2021), *I rotoli di Ercolano, la papirologia virtuale e l'edizione critica digitale dei papiri: alcune riflessioni*, in *Tracing the Same Path. Tradizione e innovazione nella papirologia ercolanese / Tradition und Fortschritt der herkulanischen Papyrologie zwischen Deutschland und Italien*, ed. by M. D'Angelo – H. Essler – F. Nicolardi, Naples, 163–7.

Reggiani, N. (2022a), *The Digital Edition of Ancient Sources as a Further Step in the Textual Transmission*, in *Digital Text Analysis of Greek and Latin sources; Methods, Tools, Perspectives*, ed. by S. Chronopoulos – F. K. Maier – A. Novokhatko, Classics@ 20, https://classics-at.chs.harvard.edu/the-digital-edition-of-ancient-sources-as-a-further-step-in-the-textual-transmission.

Reggiani, N. (2022b), *Towards a Socio-Semiotic Analysis of Greek Medical Prescriptions on Papyrus*, in *Novel Perspectives on Communication Practices in Antiquity: Towards a Historical Social-Semiotic Approach*, ed. by K. Bentein – Y. Amory, Leiden – Boston, 113–30.

Reggiani, N. (2023a), *Knowledge Construction in Progress: From Paratext to Marginal Annotations in the Greek Medical Papyri*, in *Knowledge Construction in Late Antiquity*, ed. by M. Amsler, Berlin – Boston, 133–53.

Reggiani, N. (2023b), *What Is a Book? The Ideology of Materiality in Ancient Greek and Roman Writing Technology*, in *New Approaches to the Materiality of Text in the Ancient Mediterranean. From Monuments and Buildings to Small Portable Objects*, ed. by E. Angliker – I. Bultrighini, Turnhout, 95–107.

Reggiani, N. (2024a), ed., *Materialità della medicina antica. Aspetti grafici e materiali dei papiri medici dall'antico Egitto*, Parma.

Reggiani, N. (2024b), *The Artificial Papyrologist at Work*, in *Decoding Cultural Heritage: A Critical Dissection and Taxonomy of Human Creativity through Digital Tools*, ed. by F. Moral-Andrés – E. Merino-Gomez – P. Reviriego, Berlin, in press.

Reynolds, L. D. – Wilson, N. G. (1991), *Scribes and Scholars. A Guide to the Transmission of Greek and Latin Literature*, Third Edition, Oxford.

Roberts, C. H. (1956), *Greek Literary Hands. 350 B.C. – A.D. 400*, Oxford.

Romanello, M. – Berti, M. – Boschetti, F. – Babeu, A. – Crane, G. (2009), *Rethinking Critical Editions of Fragmentary Texts by Ontologies*, in *Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies. Proceedings of 13th International Conference on Electronic Publishing (Milano, 2009)*, Milan, 155–74. [https://elpub.architexturez.net/doc/oai-elpub-id-158-elpub2009].

Roselli, A. (2012), *Galeno e la filologia del II secolo*, in *Vestigia notitiae: scritti in memoria di Michelangelo Giusta*, ed. by E. Bona – C. Lévy – G. Magnaldi, Alessandria, 63–79.

Roselli, A. (2020), *Galen's Practice of Textual Criticism*, in *From Scribal Error to Rewriting. How Ancient Texts Could and Could Not Be Changed*, ed. by A. Aejmelaeus – D. Longacre – N. Mirotadze, Göttingen, 53–72.

Sarri, A. (2018), *Material Aspects of Letter Writing in the Graeco-Roman World. 500 BC – AD 300*, Berlin – Boston.

Schubert, P. (2009), *Editing a Papyrus*, in *Oxford Handbook of Papyrology*, ed. by R. S. Bagnall, Oxford, 197–215.

Schwendner, G. W. (1988), *Literary and Non-Literary Papyri from the University of Michigan Collection*, PhD Diss., University of Michigan, Ann Arbor.

Stolk, J. V. (2015a), *Dative by Genitive Replacement in the Greek Language of the Papyri: A Diachronic Account of Case Semantics*, Journal of Greek Linguistics 15, 91–121.

Stolk, J. V. (2015b), *Case Variation in Greek Papyri. Retracing Dative Case Syncretism in the Language of the Greek Documentary Papyri and Ostraca from Egypt (300 BCE–800 CE)*, PhD Diss., University of Oslo.

Stolk, J. V. (2015c), *Scribal and Phraseological Variation in Legal Formulas: ὑπάρχω + Dative or Genitive Pronoun*, JJP 45, 255–90.

Stolk, J. V. (2018), *Encoding Linguistic Variation in Greek Documentary Papyri. The Past, Present and Future of Editorial Regularization*, in Reggiani 2018b, 119–37.

Tarte, S. (2011a), *Papyrological Investigations: Transferring Perception and Interpretation into the Digital World*, Literary and Linguistic Computing 26, 233–47. [https://doi.org/10.1093/llc/fqr010]

Tarte, S. (2011b), *Digitizing the Act of Papyrological Interpretation: Negotiating Spurious Exactitude and Genuine Uncertainty*, Literary and Linguistic Computing 26, 349–58. [https://doi.org/10.1093/llc/fqr015]

Tarte, S. (2011c), *Digital Visual Representations in Papyrology: Implications on the Nature of Digital Artefacts*, working paper, https://www.academia.edu/776645/Digital_Visual_Representations_in_Papyrology_Implications_on_the_Nature_of_Digital_Artefacts.

Tarte, S. (2012), *The Digital Existence of Words and Pictures: The Case of the Artemidorus Papyrus*, Historia 61, 325–36.

Tarte, S. (2016), *Of Features and Models: A Reflexive Account of Interdisciplinarity across Image Processing, Papyrology, and Trauma Surgery*, in *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange and Public Engagement*, ed. by G. Bodard – M. Romanello, London, 103–20.

Terras, M. [M.] (2005), *Reading the Readers: Modelling Complex Humanities Processes to Build Cognitive Systems*, Literary and Linguistic Computing 20, 41–59. [https://doi.org/10.1093/llc/fqh042]

Terras, M. M. (2006), *Image to Interpretation. An Intelligent System to Aid Historians in Reading the Vindolanda Texts*, Oxford.

Terras, M. M. (2011), *Artefacts and Errors: Acknowledging Issues of Representation in the Digital Imaging of Ancient Texts*, in *Kodikologie und Paläographie im digitalen Zeitalter 2 / Codicology and Palaeography in the Digital Age 2*, ed. by F. Fischer – C. Fritze – G. Vogeler, Norderstedt, 43–61.

Timpanaro, S. (2004), *La genesi del metodo del Lachmann*, Turin.

Totelin, L. M. V. (2009), *Galen's Use of Multiple Manuscript Copies in His Pharmacological Treatises*, in *Authorial Voices in Greco-Roman Technical Writing*, ed. by L. C. Taub – A. Doody, Trier, 81–92.

van Thiel, H. (2010), ed., *Homeri Ilias*, Hildesheim – Zürich – New York.

Vierros, M. (2018), *Linguistic Annotation of the Digital Papyrological Corpus:* Sematia, in Reggiani 2018b, 105–18.

West, M. L. (1998), ed., *Homeri Ilias, I Rhapsodias I-XII continens*, Stuttgart – Leipzig.

Willis, W. H. (1984), *The Duke Data Bank of Documentary Papyri*, in *Atti del XVII Congresso Internazionale di Papirologia*, Naples, I, 167–73.

Youtie, H. C. (1963), *The Papyrologist: Artificer of Fact*, GRBS 4, 19–32. [reprinted in *Scriptiunculae*, Amsterdam 1973, I, 9–23]

Youtie, H. C. (1966), *Text and Context in Transcribing Papyri*, GRBS 7, 251–8.

Youtie, H. C. (1974), *The Textual Criticism of Documentary Papyri. Prolegomena*, London. [1st ed. 1958]

Andrea La Veglia

Being a Classicist in the Digital Age

New Challenges and Cultural Paradigms Between Copyright Issues and Open Access

1 Introduction

> Gli umanisti, con poche eccezioni, non sembrano più essere al centro dei processi di diffusione della cultura, né come gestori, né come produttori, né come formatori.[1]

This peremptory and resigned statement, dating back to 2010, described the crisis of the *studia humanitatis*. Today, thirteen years later, after so many new innovations in the digital sphere that have led to various reformulations of the public telematic space, it is necessary to ask whether this statement is still valid. What is, then, the role and, above all, the responsibility of humanists in a new network that is no longer limited to collecting data but rather manages to relate them autonomously to each other, approaching what is commonly referred to as the *semantic web*?[2] How, then, has the humanities academic system changed in the digital age?

Throughout the history of the media, we can see that the medium which conveys cultural contents has always had a considerable influence on the content itself, because the way to share a content structurally modifies the way of thinking about that content: in Aristotelian terms, when the *matter* of an entity changes, its *form* engages a change as well. With the introduction of IT systems and Internet, the *matter* has lost its consist-

This reflection was born during a course at the Scuola Superiore Meridionale (SSM) held by Prof. Fabio Dell'Aversana about A.I. and Copyright. From those suggestions I produced a talk that I gave for the National Linux Day of October 23, 2021. The general reflection about Digital Humanities comes from my educational path at the University of Naples "Federico II": at first, I attended the Apple Developer Academy and then I graduated in Classics. For these reasons, I would like to thank all my professors and *tutores* of the SSM and all friends of NaLUG (Napoli GNU/Linux Users Group), particularly Vincenzo Palladino, who encouraged me in this research insight. I am very grateful to Maria Carla Maturo and Alessandro Russo for the formal revision of paper draft. Last but not least, a very special thank is due to Prof. Nicola Reggiani for inviting me to present this paper, as well as for the bibliographical suggestions (including one of his unpublished papers), for the references to papyrological databases he added and for the final revision. All hyperlinks last accessed on 21.7.2024.

1 Numerico – Fiormonte – Tomasi 2010, 8. ["Humanists, with few exceptions, no longer seem to be at the centre of the processes of culture dissemination, neither as managers, nor as producers, nor as educators"]

2 Tissoni 2010, 48.

https://doi.org/10.1515/9783111070162-004

ence, so that the medium has resulted invisible, and this has therefore affected the static nature of the *form*. People noticed that the transition from a real to a virtual space is not a mechanical *transfer*, but it is a real *translation* from a language to another one.[3] From this perspective, the main problem that philologists face in publishing their works on the web is related to the ontology of the work itself. Finally, it is necessary to understand not only how IT could help Humanities and how Humanities could improve IT, but what are the new paradigms that can help to complement the two fields of knowledge each other, according to the notion of "cultural informatics".[4] So, how has the humanities academic system changed in the digital age? What are the new pathways that could be followed in facing all these new challenges?

The progression of such an argument can be described as elliptical, as it gravitates around two *foci*, linked by an intrinsic complementary relationship. The two *foci* in question are: the role of information technology as *ancilla humanitatum* and, conversely, the role of the humanities as *ancillae technologiae*, in developing open-source models.

In this way, we could attempt to understand what role the humanist has assumed after the telematics revolution, particularly in the guise of classicist and philologist, analyzing the obstacles (s)he encounters in defining her/his status and the technological and legislative tools that instead allow for her/his legitimization.

2 The 'technological leap' in the history of the digital humanities

The set of activities in which the humanist is at the centre of digital processes and dynamics is referred to as *Digital Humanities* (DH) and these activities clearly include Digital Papyrology. Recently, in defining what DHs are, J. Drucker has written that they operate in the "intersection of computational methods and humanities materials"[5] and in so doing, he has assigned the new technologies a methodological role and has defined humanities as an object of study. This definition is important in order to contextualize the problems we are going to analyze. At first reading, it would seem to contrast with the idea of complementarity introduced earlier. However, it should be noted that by "methods" J. Drucker refers to both practical tools, i.e. the utilization of tools, and theoretical approaches, i.e. computational thinking. His definition consequently sums up the two trends of digital humanists: the first consisting in using web-based tools and software for the creation and dissemination of humanities contents, and the second is the

3 Fiormonte 2003, 9.

4 Crane – Bamman – Jones 2013, 52–5.

5 Drucker 2021, 1.

utilization of computational logics for the modelling of new communicative paradigms. In the latter case, the new technologies are an object of study of the humanities.

The new phase of DH seems to focus on overcoming the previous theses and antitheses, reaching a synthesis. In this synthesis, the grammatical subordination that subsists between 'digital' and 'humanities' would not imply a theoretical subordination, even less an inverse subordination, as the now obsolete syntagma 'Humanities Computing' would suggest,[6] but precisely a complementarity.

A path in this direction seems to have been mapped out as early as 2008, by a collective of humanists at the University of California, Los Angeles (UCLA), who authored the famous *Digital Humanities Manifesto*,[7] and who later emphasized that "the mere use of digital tools for the purpose of humanities research and communication does not qualify as Digital Humanities".[8] In their epistemological analysis, they define DH as the possible way to redraw the boundaries of the digital world that apparently excludes humanists. In this regard, it should be noted that while DH were already the subject of epistemological analysis in 2008, it can be rightly said that such a field is not only an independent discipline in its own right, but also a well-established one, and thus the skepticism shown by some scholars in the scientific community towards defining DH as an independent field seems at least anachronistic. In this way, therefore, they defined DH by its objective (*causa finalis*):

> The Digital Humanities seeks[9] to play an inaugural role with respect to a world in which, no longer the sole producers, stewards, and disseminators of knowledge or culture, universities are called upon to shape natively digital models of scholarly discourse for the newly emerging public spheres of the present era (the www, the blogosphere, digital libraries, etc.), to model excellence and innovation in these domains, and to facilitate the formation of networks of knowledge production, exchange, and dissemination that are, at once, global and local.

This call to action thus argues that using the telematic space to produce new culture in the humanities implies a rethinking of criteria for organizing the structures producing that culture. The first thing that needs to be asked is which points of continuity and points of discontinuity there are in the transition from a *traditional* to a *digital* approach within the *humanities*.

6 Cf. Scholes – Wulfman 2008.

7 Drucker – Lunenfeld – Presner – Schnapp 2008. The manifesto was written of the UCLA seminar entitled *What Is(n't) Digital Humanities?* held during the 2008-09 academic year and it was posted on the *UCLA-Digital Humanities & Media Studies* website. In 2009, T. Presner and J. Schnapp uploaded a PDF file on their blogs that reworks the contents of the manifesto using an educational language, thus naming it *Manifesto 2.0* (http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf).

8 Cf. Burdick – Drucker – Lunenfeld – Presner – Schnapp 2012, 122.

9 The syntagma 'digital humanities' is generally singular when referring to the field of study and plural when referring to 'diversity of practices'. Cf. Drucker 2021, 17.

Indeed, a continuity factor is undeniable: the digitization of the book constitutes the latest 'technological leap'[10] in the history of the transmission of the written text and it should be regarded as the natural continuation of such an unstoppable long-term process.[11] However, it is equally undeniable that the transition from the paper medium to the virtual one has caused a "paradigm shift", which, according to Kuhn, occurs with the advent of every scientific revolution, that is, when a new theoretical discovery generates the crisis of an entire system of thought.[12] The shift from printed paper to the electronic medium, therefore, can be related in importance to the shift from the scroll to the codex or even the shift from orality to writing.

Considerable insight can be taken from DH first project:[13] Father Roberto Busa's *Index Thomisticus*, which – as is well known – is a concordance of *lemmata* of the entire corpus of Thomas Aquinas' writings. Father Busa had the insight to use the I.B.M. punching machine to automatically create the concordance sheets of his *Index*, from which he had then selected the portions of the text that had to be printed. In this very early phase of DH, or rather, of Italian *Informatica Umanistica*, technology played a role in the realization phase of the editorial product, as a support to the editor, but the final product to be published was still the printed volume. On the other hand, it was impossible, at that time, to imagine a distribution on electronic media, due to the cognitive and economic limitations that prevented even the conception of the idea of 'electronic publication'.

In fact, the idea of the electronic edition of a philological work could only be conceived after another significant technological leap: namely, the market launch of the first *Macintosh* in 1984. The interface with icons and windows still used in contemporary personal computers revolutionized the relationship between man and machine, and specifically, the development of the first word processing programs directed the humanists toward new goals and methodologies. Computers were no longer just the tool for storing and managing huge amounts of data, but became in *potentiality* laboratories where texts and databases could be consulted through more accessible software and interfaces, even "col risultato di focalizzare l'attenzione sul risultato visibile, e non sulle procedure sperimentali."[14]

Thus, the first DH project that published texts on CD-ROM was the TLG[15] and it was followed a few years later by the *Index Thomisticus*.[16] Therefore, we can say that the *Index Thomisticus* has followed in its own history the evolution of the technological tools and of the relationship between new technology and the humanities. In 2004, Father Busa him-

10 For the concept of 'technological leap' cf. Milanese 2020, 22–34.

11 For the first formulation of the long-term concept see Braudel 1958.

12 Kuhn 1970, 77–91.

13 Hockey 2004, 4; Milanese 2020, 46.

14 Orlandi 2012, 51. ["with the result of shifting focus to the visible output, and not to the experimental procedures"]

15 Pantelia 2000, 3.

16 Cf. Busa 1992.

self said that he was surprised by the enormous technological develop-ments that had taken place from the beginning to the conclusion of his project ("digitus Dei est hic!"[17]), which was further updated during the following year when a website was activated.

The TLG and the *Index Thomisticus* were forerunners for many other text indexing and digitization projects that survive to this day: notable for their longevity, among others, are the *Dartmouth Dante Project* (DDP)[18] for the Divine Comedy and the *Packard Humanities Institute*'s (PHI) *Classic Latin Text*[19] for Latin classics, as well as the PHI#7 *Duke Documentary Papyri*, subsequently flown into the *Duke Databank of Documentary Papyri* and now in *Papyri.info*.[20] These projects, just like the *Index Thomisticus*, were also initially meant for print publications and then were updated to be distributed starting from the 1980s as text files and AutoPlay on CD-ROM and finally on the Web towards the end of the century.[21]

Concurrently with the development of these projects, epistemological studies started that sought a dialogue between computer science and the humanities, and in 1966 the first journal exploring the intersection between the two fields of study, namely *Computers and the Humanities,* was born. In the early 1970s, moreover, the most important European centres in the field were founded, including those of *Informatica Umanistica* at the University of Rome "La Sapienza" and at the University of Pisa.[22]

In 1987, the *Text Encoding Initiative* (TEI) was established after a conference organized by the *Association for Computers and the Humanities* (ACH),[23] aimed at finding an interoperability standard for transcribing, formatting and encoding humanities texts. This underscores the need to foster initiatives of a collaborative nature in an inter-university perspective, which the free access to Internet cleared shortly thereafter was making possible.

At the current state of things, in the field of DH, projects have increased exponentially[24] and multiple sub-disciplines within the set of DH have been defined. Among these we are going to analyze here Digital Philology, understood here in the sense of *Textual Criticism*,[25] by calque of the Italian *Filologia*.[26]

17 Busa 2004, XVI. Cf. also Busa 1980.

18 https://dante.dartmouth.edu/about.php.

19 https://latin.packhum.org.

20 See Reggiani 2017, 210–31.

21 Cf. Pantelia 2000 and the history section of http://stephanus.tlg.uci.edu/history.php. The PHI *Duke Documentary Papyri* database was conceived on digital medium from its very beginnings, but an earlier project started at the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) at the University of Liège, then discontinued, intended to produce printed papyrological concordances, indexes, and text editions from a digitized database: see Reggiani 2017, 207–9.

22 Cf. in this regard Ciotti 2018 and Orlandi 2012, last updated in Orlandi – Tomasi 2023.

23 https://ach.org.

24 For a timeline of DH projects see https://www.historyofinformation.com/index.php?cat=68.

25 Cf. Reeve 1999.

26 Varvaro 2012, 11 ff.

3 The model of 'electronic mirror image'

In the program of the *Thesaurus Linguae Graecae* (TLG),[27] the digitization of the corpus of Greek literature consists of choosing a critical edition for each text and then creating "an electronic mirror image of the source edition from which it derives".[28] F. Tissoni notes that for the purposes of intertextuality work, it is not sufficient to create an electronic mirror image of each edition, but it is necessary to work on the text in order to normalize the word-forms according to a criterion of consistency, as for example it was done for the Library of Latin Texts.[29] As previously mentioned, in the transition from print media to digital space we need to *translate* content into a new language, by taking full advantage of all the features of the new media and by making skillful use of the potential of hypertext and mark-up languages.

However, there seems to be a conceptual brake in this translation operation especially on the part of classicists. As F. Tissoni observes, in fact, "poiché la moderna filologia è nata quando la stampa già esisteva, sembra che la forma-libro sia non solo naturale, ma anche l'unica possibile per ospitare una edizione critica."[30]

Critical edition cannot be separated from the book-form, even though this does not benefit the print media market. Indeed, there is a widespread attitude in academia to violate copyright altogether by downloading print proofs of books and scholarly articles via platforms such as *Library Genesis* or *ZLibrary* and *Sci-Hub* that continue to survive despite regulatory measures trying to hinder them.[31]

Confirming the persistence of the book-form in the digital world, moreover, it can be noted that scientific articles and conference proceedings published online are often distributed in Portable Document Format (PDF), thus designed for printing or at least for digital consultation that resembles as closely as possible the consultation of a paper document: infrequently these kinds of scientific texts are published as hypertexts.[32]

At this moment, therefore, both print and digital publishing are at an *impasse*, as the former is in crisis due to piracy, but at the same time the latter does not seem to

27 For a history of the TLG see Pantelia 2003.

28 Pantelia 2000, 2; see also Brunner 1991, 63.

29 Tissoni 2015, §§ 29–33.

30 Tissoni 2010, 120. ["Since new textual criticism was born when printing already existed, it seems that the book-form is not only natural, but also the only possible one to accommodate a critical edition"] F. Tissoni here writes *moderna filologia* expressing a different meaning from *filologia moderna*, which in Italian means "textual criticism of modern works", as intended in the title of J. McGann's book named "A critique of modern textual criticism." So, I avoid to use the adjective 'modern', but at the same time I do not want to refer to the "New Philology." Cf. McGann 1992; Varvaro 1999b.

31 The most recent measure was the FBI indictment that led to the arrest of the two website managers of *ZLibrary*, see FBI, *Two Russian Nationals Charged with Running Massive E-Book Piracy Website* (press release), November 16, 2022 (https://www.justice.gov/usao-edny/pr/two-russian-nationals-charged-running-massive-ebook-piracy-website).

32 E.g., just consider the way content is published on Open Edition (https://www.openedition.org).

have achieved a fair degree of loyalty with readers. It can be said that material culture, paradoxically, is at the centre of digital dynamics without benefiting from them.

On a semiotic analysis, it can be observed that it is difficult to place a 'specific' of the print media, the critical edition, in the virtual space, where the 'specifics' are *blogs* and *wiki* resources.[33]

Clearly, there is a problem of perspective. It is the *substance* of the medium that must affect the *form* of the content and the opposite cannot happen. Committing the naïve anachronism of making the Internet an "electronic mirror image" of reality would be tantamount to persisting in using a fountain pen to write on a computer screen. By doing so, we inadvertently fall into the trap of adhering to what McLuhan[34] referred to as "rear-view mirror" logic, according to which people constrains novelty within pre-existing paradigms.

So, the digitization of a text – even when it is based on an existing philological work – is equivalent to a re-edition work for all intents and purposes, and for this reason it must be performed by philologists and cannot be delegated to other figures.

The issue is the scarcity of digital philologists. As early as twenty years ago, J. McGann observed that a "network of digital storage, access, and dissemination" of humanities was being established, but he regretted that few of those involved were trained philologists.[35] As we have read at the beginning of the paper, this issue is found again in 2012 and is still evident today, unfortunately, more than ten years later.

Many philologists are, alas, wary of the Web because of both a general reluctance to new technologies, since the philologist is proverbially fond of printed paper, and a view of the Internet as a place where copyright is not protected. The fear is not entirely unfounded, but the first step to overcome it involves awareness of the dangers and knowledge of the tools to counter it.

33 'Specific' is used here as a noun, in the Italian meaning of the term, indicating a "singular, typical element that distinguishes and characterizes a given context". In this regard, I would like to thank Prof. Renata Bellucci for formulating this concept in semiotics.

34 McLuhan – Fiore – Agel 1996, 100 and Milanese 2020, 26–8.

35 McGann 2002.

4 A classicist's journey between copyright, public domain and fair use

> Practice digital anarchy by creatively undermining copyright and mashing up media.[36]

In 1993 the CERN decided to put Tim Berners Lee's World Wide Web software[37] into the public domain, thus making access to the Internet free of charge. Since then, the Web has become the free space *par excellence* and the idea was born that it should include only free content.

So, it was not a coincidence that as early as 1999 *Napster* was launched. It was an Internet Content Provider (ICP) that achieved a dual record: it was the first free peer-to-peer content sharing system, but at the same time the first platform to carry pirated music. Precisely for copyright violation it was sued by the owners of the shared music and closed its doors after only two years from its launch.[38]

Clearly, the end of *Napster* did not spell the end of software piracy, in fact, to give an example in the context of academic publications, "the new Napster" is now *Sci-Hub* for its record of illegal downloads.[39]

One could briefly say that the telematic *agorà*, because of its very nature, amplifies that process of 'deregulation' taking place in real society according to Bauman, who in his famous theory called it "liquid" because it is the result of the processes of melting down the solid structures that held it up with the aim of freezing into "new and improved solids", but "the task of constructing a new and better order to replace the old and defective one is not presently on the agenda".[40]

In the same way, then, it is observed that in the transition from paper to electronic media for text preservation, authorship also lost its solidity and seems to have dissolved into a 'gradient', the same as Alberto Varvaro observed in the Middle Ages (*gradiente di autorialità*) between texts with high authorship (classical and sacred texts) and texts with low authorship (chronicles and apocryphal texts). Low-authorship texts were freely reworked by copyists, as they were perceived as texts without owner, and today we might say copyright-free.[41] Using the same categories, it could be said that people

36 Drucker – Lunenfeld – Presner – Schnapp 2008, § 7.

37 At the following link you can read the document that officially put the World Wide Web into the public domain on April 30, 1993: https://cds.cern.ch/record/1164399?ln=it. The CERN uploaded also a reconstruction of the first website: http://info.cern.ch/hypertext/WWW/TheProject.html.

38 For an history of Napster see *Encyclopædia Britannica Online s.v.* [https://www.britannica.com/topic/napster].

39 González-Solar – Fernández-Marcial 2019.

40 Bauman 2000, 2–6.

41 Varvaro 1999a, 402 and 420.

perceive books and articles published in print as they had higher level of authorship than material published online.

This is due to the fact that the web is considered, just as in its early days, the space of amateurism and of free contents, but mainly in the sense of free-of-cost rather than for freedom of expression, as G. Lovink and N. Carr note.[42] Just for this reason R. Mordenti considers all attempts to charge for music and book files distributed on the web useless, defiantly calling them *non libri* and *non dischi*.[43]

And this is also the case in the field of Philology, where digital critical editions are regarded as non-products: when they are consulted, they are not even cited, because the norm is to cite the related printed edition. Of course, if the text has only been edited digitally, one cannot do otherwise, but generally citing the electronic critical edition available online, where it is not strictly necessary, is perceived as a *deminutio*.

Is there any protection for the philologist when (s)he publishes her/his work on the Web? To answer this question, we must first delve into what the Intellectual Property (IP) protection of the *classicist-editor* is in print media.

T. Margoni and M. Perry note that "the protection of scientific and critical editions is not present in the Berne Convention, TRIPS, or any WIPO Treaty"[44], i.e. the main international treaties concerning copyright.[45] Consequently, the treatment of IP rights concerning critical editions remains contingent upon the distinct legal systems of individual nations. We are to going to have a look at American legislation and European directives.

– Since they are not mentioned in the Copyright Act of 1976,[46] critical editions in the U.S. fall within the framework of derivative works and, as in all derivative works, copyright protects only new and significant contributions to the original text, thus in the case of scholarly editions, mainly the introduction, critical apparatus and commentary,[47] but in the case of Dead Sea Scrolls the philologist was able to copyright the text itself.[48] This copyright has the same duration as the copyright of an original work, i.e. up to 70 years after the author's death, unless otherwise stated in the publishing contract.

42 Lovink 2008, XIII–V.

43 Mordenti 2012, 176. ["non-books", "non-disks"]

44 Margoni – Perry 2011. I thank Prof. Mark Perry (University of New England) for suggesting some research insight in this way.

45 The authors refer to the *Berne Convention for the Protection of Literary and Artistic Works* of 1886, amended in 1979 [URL:https://www.wipo.int/wipolex/en/text/283698], to the *Trade Related Aspects of Intellectual Property* (TRIPS) of 1994, amended in 2017 [URL: https://www.wto.org/english/docs_e/legal_e/31bis_trips_01_e.htm], and to the *World Intellectual Property Organization Copyright Treaty* of 1996 [URL: https://www.wipo.int/wipolex/en/treaties/textdetails/12740].

46 Codified now in Title 17 of the USA Code and available at https://www.copyright.gov/title17/title17.pdf.

47 Cf. See Margoni – Perry 2011, 165, where they address the issue from a comparative law perspective.

48 For an overview on this case see Nimmer 2001 and Oakes 2001 and cf. note 76. Special thanks are due to Dr. Bohdan Widła, who suggested to delve into this interesting episode.

– In Europe, on the other hand, since *Council Directive 93/98/EEC (October 29, 1993), harmonizing the term of protection of copyright and certain related rights*,[49] critical editions have had a specific framework, within the framework of neighbouring or related rights, i.e. the rights of a work related to the author's work, just as the right of performance. According to Article 5[1] of this directive, in fact, "Member States may protect critical and scientific publications of works which have come into the public domain". This directive is most likely inspired by the German Copyright Act of 1965,[48] where it is stated, in Article 70, according to the last amendment:

> Ausgaben urheberrechtlich nicht geschützter Werke oder Texte werden in entsprechender Anwendung der Vorschriften des Teils 1 geschützt,[50] wenn sie das Ergebnis wissenschaftlich sichtender Tätigkeit darstellen und sich wesentlich von den bisher bekannten Ausgaben der Werke oder Texte unterscheiden.[51]

Moreover, in the law, the critical edition as a whole is protected as long as it has significantly altered the text. It is no accident that critical edition is mentioned for the first time in the regulatory system of Germany, which is the home of philology. The philologist's right, as a related right, has a shorter term than the 70 years of copyright and is even shorter than the 50 years established for performance rights: Article 5[2] of the European Term Directive indicates a maximum term of protection of 30 years from publication. This term is also inspired by the German law that in 1965 provided it for 10 years and, currently, for harmonization with the European Directive, for 25 years from publication or (in the absence of publication) from the production of the text.

Following the publication of this harmonization directive – which, according to T. Margoni and M. Perry, had more the purpose of 'dis-harmonization' – member states of the European Union have introduced in their respective Copyright Acts a reference to scholarly editions. The Italian Copyright Law[52] inserted in 1997 a reference to scientific editions with Articles 85-quater and 85-quinquies, where it is pres-

49 URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A31993L0098. The directive was replaced by *Directive 2006/116/EC of the European Parliament and of the Council (12 December 2006) on the term of protection of copyright and certain related rights* (https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32006L0116).

50 *Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz)*, Federal Law Gazette I, 9/09/1965, p. 1273) [https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl165s1273.pdf#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl165s1273.pdf%27%5D__1694209054243] last amended by Article 25 of the Act of 23 June 2021 (Federal Law Gazette I, p. 1858) [https://www.gesetzeiminternet.de/englisch_urhg/index.html]. Cf. Margoni – Perry 2011, 166.

51 ["Editions of works or texts which are not protected by copyright shall be protected *mutatis mutandis* under the provisions of Part 1 if they represent the result of scientifically organized activity and differ substantially from previously known editions of the works or texts" (official translation from https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html]

52 L. 633/1941 *Protezione del diritto d'autore e di altri diritti connessi al suo esercizio*, arts. 85-quater e 85-quinquies.

cribed that the editor (in Italian, *editore* or *curatore*) is the owner of the rights of the work and divides according to contractual indications the economic exploitation rights (*diritti di utilizzazione economica*) with the publisher (in Italian, *editore* or *casa editrice*),[53] but in any case, they have the right to the indication of the name. These economic exploitation rights last for 20 years, below the European average, which is 25 years.

Digital critical editions can be either (1) the result of a new collation and interpretation of the *testimonia*, (2) the result of collation of two or more authoritative critical editions,[54] or (3) the result of digitization of a single critical edition with or without intervention by the digital editor.[55] In all of these cases the creative work of the philologist is present to a lesser or greater extent, and in cases 2 and 3 it is mixed with the creative work of the philologist of a paper edition, whose rights still subsist if the years stipulated by the jurisdiction of the country in question have not passed (a very wide time range, varying from 20 to 70+ years!).

The digital philologist (or the academic institution or publisher) therefore has to pay royalties to the philologist of the print edition who still enjoys copyright, but there are some exceptions. In the U.S.A., digital philologists can avail themselves of Fair Use, a legal doctrine defined by paragraph 107 of the U.S. Copyright Act. According to this paragraph, we can freely use the copyrighted material for purposes related to the dissemination of culture, given the nature of the copyrighted work, the amount of the portion used of copyrighted work, and the effect of the use upon the potential market, a doctrine connected with *Free Use* provided for in Articles 10 and 10-bis of WIPO Lex and, for example, incorporated into the Italian Copyright Law in 2008.[56]

The roots of this copyright exemption can be found in the U.S. constitutional charter, in which one can read the IP Clause[57] according to which the government's objective is "[t]o promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."

N. Wiener observes that this kind of restriction is closely linked to the American culture of the 18th century, according to which it was difficult to claim IP over the discovery of a law of nature, as could be the patenting of a machine, as it happens "in many

53 Mordenti 2012, 171 and 179.

54 An example could be the *Bibbia Edu* project, Fondazione di Religione Santi Francesco d'Assisi e Caterina da Siena and Conferenza episcopale Italiana (CEI), Rome 2008-19 (bibbiaedu.it).

55 An overview on digital critical editions was made by Magnani 2018, 86–90.

56 Art. 2 of L. 2/2008, *Disposizioni concernenti la Società italiana degli autori ed editori* (G. U. n. 21, 25/01/2008) that provides the last emendation to art. 70 of L. 633/1941. I had an interesting email correspondence with Bohdan Widła about this: there is no real application of 'fair use' in European countries, but it is often used to refer to the idea behind this American legal doctrine. Because of this I have used 'fair use' in inverted commas infra in reference to an Italian website.

57 *Constitution of the United States*, 15/09/1787, art. 1, sec. 8[8] [https://www.senate.gov/about/origins-foundations/senate-and-constitution/constitution.htm]

other countries with similar industrial practices".[58] Thus, if in the American Constitutional Charter the principle of freedom of scientific works was an expression of utilitarianism, today it can be re-interpreted in function of easier dissemination of culture in the digital age.

According to Fair Use in U.S. and to copyright exceptions implemented in European countries,[59] the philologist may integrate into her/his projects part of copyrighted printed critical edition within the limitations of the law in force in the country of publication of that edition, but more importantly (s)he may add an open license to her/his edition, according to the conditions we are going to see later.

But let us start with an example, namely, DH project *digilibLT*,[60] developed by the University of Eastern Piedmont, which has digitized the texts of the most authoritative critical editions of Late Latin works and organized them so that each portion of text can be traced back to the page of the reference edition. When the user copies the bibliographical reference of the edition, at the moment of pasting it into her/his word processor (s)he discovers that (s)he has pasted – next to the very bibliographical reference – a text string that was invisible on the website (Figs. 1–2):

> Text distributed under a Creative Commons Attribution, Noncommercial, Share Alike 3.0 Italy license. Therefore, please cite the passage, indicating that it comes from the digilibLT project website, https://digiliblt.uniupo.it, and indicate the names of the people who worked on the digital edition of the cited work.

An initial reflection can be made on this ploy. The fact of remembering to cite the critical edition registers the digital philologist's legitimate fear of not having his or her rights recognized because digital critical editions are not considered authoritative.

An opposite statement was written by the editors of the Latin Library, who recall that their project does not aim to replace a critical edition and reiterate that all scanned and formatted critical editions are in the Public Domain (PD) in order to protect themselves (Fig. 3).

The editors of the Classical Latin Texts, who have also used critical editions not yet in the PD, prevent readers from accessing the texts they have digitized without first declaring that they access to edited contents for personal use only and in accordance with the principles of Fair Use. This declaration is required for consultation of all other PHI's digital text archives (Fig. 4).

58 Wiener 1989, 114.

59 Angelopoulos 2022.

60 Site link: https://digiliblt.uniupo.it, see Lana 2012.

Fig. 1: The sample source text at DigilibLT (https://digiliblt.uniupo.it/opera.php?gruppo=opere&iniziale=all&id=DLT000616).

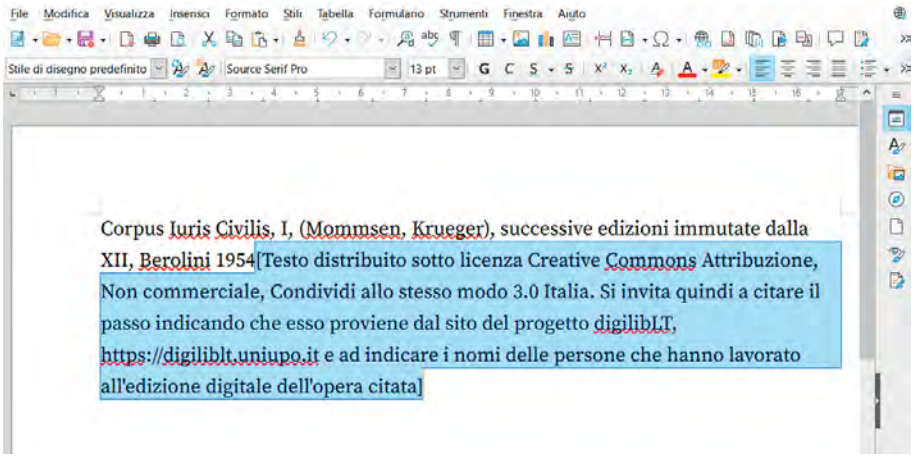
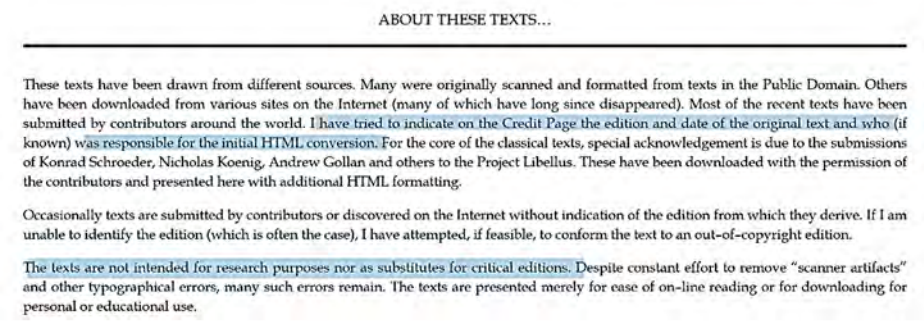


Fig. 2: The pasted text.



ABOUT THESE TEXTS...

These texts have been drawn from different sources. Many were originally scanned and formatted from texts in the Public Domain. Others have been downloaded from various sites on the Internet (many of which have long since disappeared). Most of the recent texts have been submitted by contributors around the world. I have tried to indicate on the Credit Page the edition and date of the original text and who (if known) was responsible for the initial HTML conversion. For the core of the classical texts, special acknowledgement is due to the submissions of Konrad Schroeder, Nicholas Koenig, Andrew Gollan and others to the Project Libellus. These have been downloaded with the permission of the contributors and presented here with additional HTML formatting.

Occasionally texts are submitted by contributors or discovered on the Internet without indication of the edition from which they derive. If I am unable to identify the edition (which is often the case), I have attempted, if feasible, to conform the text to an out-of-copyright edition.

The texts are not intended for research purposes nor as substitutes for critical editions. Despite constant effort to remove "scanner artifacts" and other typographical errors, many such errors remain. The texts are presented merely for ease of on-line reading or for downloading for personal or educational use.

Fig. 3: The discussed statement at The Latin Library (www.thelatinlibrary.com/about.html).

Classical Latin Texts

A Resource Prepared by
The Packard Humanities Institute

I agree to use this web site only for personal study and not to make copies except for my personal use under "Fair Use" principles of Copyright law.

Click here if you agree to this License

Questions and comments can be directed to
The Packard Humanities Institute

Fig. 4: The discussed statement at Classical Latin Texts (https://latin.packhum.org).

The editors of the *Musisque Deoque* project likewise declare that both the copyrights of the critical editions they consulted and the copyrights of the digital publishers are reserved, with only exception of "Fair Use" (Fig. 5):

> All rights to the texts with apparatus contained in www.mqdq.it are reserved by the units of the *Progetto di Ricerca di Interesse Nazionale Musisque Deoque*, the editors of the work and the original authors of the documents. No use for commercial purposes is permitted without prior agreement. Reproduction and circulation in hard copy or portable electronic media (off-line) for scientific, educational or documentary use only is permitted, provided that the documents are not substantially altered in any way, and in particular retain the correct date, authorship and original source information (citation). Links from other websites are welcome, especially if notice is given to the editors (info@lutessa.it.it), to facilitate prompt notification of any subsequent changes. Any kind of mirroring on other sites, or automatic capturing of texts, is prohibited, unless specifically agreed with the editorial team.

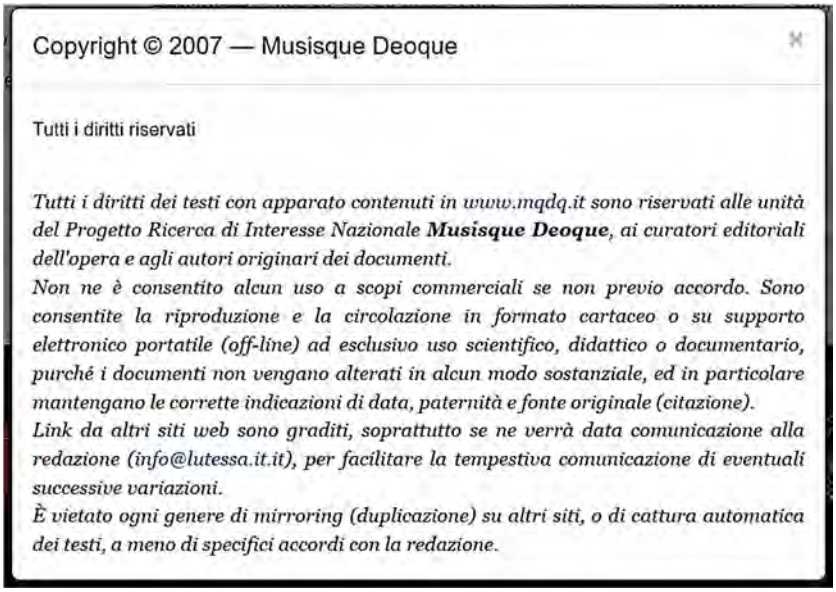


Fig. 5: The discussed statement at Musisque Deoque (https://www.mqdq.it/public).

What one might ask at this point is how to overcome the limitations that copyright poses in order to achieve a more coherent *synolus* between form-edition and digital substance, while protecting the traditional philologist and the digital philologist.

In his first volume on Digital Papyrology, N. Reggiani called for a re-interpretation of copyright and IP, proposing to imagine digitization not as a "safekeeping or copying affair", but rather "as a place where true scholarship can primarily, if not exclusively, take place," glimpsing the signs of "promising developments" in the future.[61] Certainly, the promising developments could be seen in the gradual improvement of Open Access.

5 What is Open Access? About the importance of being free, but not like a beer

The year before *Napster*'s closure (1998), Christine Peterson uttered the syntagma *open source* for the first time at a conference in Palo Alto, California. Shortly afterwards, the *Open Source Initiative* (OSI) was born, with the aim of promoting the idea of *Open Source*, which does not only mean having access to the source code, but also allowing

61 Reggiani 2017, 177.

any programmer to *fork* it, i.e. reuse it in the development of a new one.[62] The new software is obliged to inherit the *Open Source* license, otherwise the programmer incurs a criminal penalty for violation of the license in question, as stipulated in 2008 by a ruling of the United States Court of Appeals for the Federal Circuit.[63]

It is the principle of "share and share alike" that is also characteristic of the *Free Software* movement started as early as 1985 by the far-sighted action of the Massachusetts researcher Richard Stallman. The name is different because "different words convey different ideas":[64] the adjective 'free' gives philosophical substance to the logic of *open source*, since the 'practical' freedom to use a software and access the source code derives from the 'theoretical' freedom of human beings in the technological world. And Stallman points out that in the expression 'free software', the adjective 'free' is not to be understood as in 'free beer'. Rather, it is legal freedom close to 'freedom of speech'.[65]

As mentioned earlier, it is from the characteristic of the medium that the idea of content has developed. Thus, it was an almost natural process to apply the paradigm of dissemination and protection of IP of software to cultural content. In fact, in 2001, at the behest of Laurence Lessig, the *Creative Commons* (CC) organization was founded, which seemed to give a legal framework to the concept of "authorship's gradient" mentioned earlier. In fact, the CC stands between the PD and the traditional *copyright-all rights reserved* in a gradation of four licenses that reserve, in a progressive measure, only 'some rights' according to the wording of each license.

This could certainly be the solution to face the difficulty of humanities in finding their place in the digital world. The encounter between new technologies and humanities generated a wide-ranging reflection that later led to the promotion of 'Open Access' (OA), which means the "free and unrestricted online availability" of scientific research results, according to the Budapest Declaration of 2002,[66] followed by the Bethesda[67] and Berlin[68] Declarations of the following year, signed by the world's leading universities to promote open access to knowledge.[69] These statements merely reiterate the mission conceived for the World Wide Web at its beginnings in 1990: to connect research centres for the advancement of scientific knowledge in a net where Universities are the nodes.[70]

Thus, one can see why new paradigms are needed to convey cultural content on the Web. A focal point was the case of Aaron Swartz, *enfant prodige* of computer science as

62 See *History of the OSI* (2006-2018) on the history section of the official *Open Source Initiative* website (https://opensource.org/history).

63 *Jacobsen v. Katzer*, 535 F.3d 1373 (Fed Cir. 2008) [https://cafc.uscourts.gov/opinions-orders/08-1001.pdf].

64 Stallman 2009, 31.

65 Stallman 1985.

66 https://www.budapestopenaccessinitiative.org/read.

67 http://legacy.earlham.edu/~peters/fos/bethesda.htm.

68 https://openaccess.mpg.de/Berlin-Declaration.

69 D'Andrea – Toccoli 2010, 47–8.

70 Berners-Lee 1990, 14.

well as OSI and OA activist, who was investigated by the Massachusetts District Court for 7 violations with a potential maximum sentence of 35 years in prison for illegally downloading, bypassing the paywall, 5 million scholarly articles from the *JSTOR* platform.[71] The indictment led to his depression and suicide in 2013. L. Lessig, the founder of the Creative Commons license, was Swartz's lecturer and professor and gave a *lectio magistralis* in his honor at Harvard Law School. He, in moving words, called Swartz's action an "act of citizenship",[72] later writing that, in his view, the culprit in the situation was a "system of copyright built for the physical world, a system now struggling to catch up with the digital".[73]

It should be noted that the *JSTOR* platform was born precisely with the goal of fostering access to culture, and that in September 2011[74] it launched the Early Journal Content service available to non-registered users, which made articles published before 1924 in the United States and before 1876 in other countries available, thus pushing towards OA. However, the platform reiterated in 2017 that, according to the editors, it is not true "that just because something is in the public domain, it can always be provided for free", because of the cost of digitization, thus demanding its own copyright on works in the PD despite the fact that it was a minimal creative contribution (only scanning).[75] Despite this claim, in the following years, the repository of freely accessible journals for independent researchers not registered on the platform continued to grow.

There is no doubt, however, that the Aaron Swartz episode has stirred something up and raised the bar a little higher towards an open science model for the humanities.

6 Toward the Open Digital Humanities

In order to achieve an *open* dimension of the *studia humanitatis*, it is first of all necessary to recognize the technical-scientific character of the critical edition, which is research evidence on a par with the results of a scientific experiment.[76] Logical conse-

71 Superseding Indictment at 18, *United States v. Swartz*, No. 1:11-CR10260-NMG (D. Mass. Sept. 9, 2012) [https://archive.org/details/UsaV.AaronSwartz-CriminalDocument53/page/n17/mode/2up].

72 Lessig 2013a.

73 Lessig 2013b.

74 *68,000 additional free articles added to Early Journal Content* [https://about.jstor.org/news/68000-additional-free-articles-added-to-early-journal-content/].

75 *FAQ – Why not make any and all public domain content freely available?* [https://web.archive.org/web/20170511080512/http:/about.jstor.org/individuals-faq#Why_not_make_any_and_all_public_domain_content_freely_available].

76 Elkin-Koren 2001 compares deciphering of human genome with the reconstruction of Dead Sea Scrolls and notes that in the first case President and Prime Minister of US declared that the human genome belongs to all, instead in the second case the Israeli Supreme Court granted the editor Qimron the copyright of his deciphered text. The author notes that it is a paradox and that the Court confuses Arts and Science, eventually concluding that the copyright on Dead Sea Scrolls will cause problems to

quence is the right of every humanist to *open access* to the most up-to-date critical edition of a work, for the advancement of scientific progress. In part, this right is granted by the possibility of physically or digitally borrowing the text in a library, either through a document delivery service or through digital borrowing at *Archive.org*,[77] where the Teubner volumes and the Editions du Cerf whose publisher's rights have expired can be found. A mention should be made in this regard of the digital Loeb Classical Library,[78] the Oxford Scholarly Edition Online[79] and the digitization (*scil.* 'scanning') of some volumes of the Belles Lettres through Open Edition.

N. Reggiani noted that during COVID-19 pandemics and thus at a time when it was impossible to reach a library, online resources were a lifeline for papyrologists, who by virtue of *amicitia papyrologorum*, through the papyrological mailing list (Papy-list) exchanged information about the volumes available online and noticed that some paywalls had been exceptionally released during the Covid emergency.[80] This is the case of the *JSTOR* platform that to support researchers "during this challenging time in which many are unable to get to physical libraries" expanded the free read-online access to 100 articles per month[81] (Fig. 6), offer ended on June 30, 2023.[82]

The problem is clearly to find sustainable business models to allow open access to humanities content, but a first step forward would be to approach the model used by STEM: "le scienze dure (a diffusione internazionale) si rivolgono ormai da tempo agli archivi disciplinari (Pubmedcentral, Arxiv ecc.), le scienze umane (di solito a diffusione nazionale) sono invece un po' ferme."[83]

R. Mordenti emphasizes the trend in the so-called 'hard sciences' of making available not only the results of experiments, but also the insights and hypotheses of research and hopes for a future in which similarly a critical edition is made available at every stage of its creation and does it free of charge. He also proposes a reformulation of the critical edition's economic sustainability model: not *royalties,* but funding from the cultural institution from which the research that led to the critical edition originated.[84].

"viability" of future research. The Digital Scholarly Editions Manifesto of 2022 underlines the technical-scientific character of critical editions. Cf. Ciotti – Corradini – Cugliana *et al.* 2022.

77 *Borrowing From The Lending Library – Internet Archive Help Center* [https://help.archive.org/help/borrowing-from-the-lending-library/].

78 See https://www.loebclassics.com/page/history.

79 See https://www.oxfordscholarlyeditions.com/page/146.

80 Cf. Reggiani 2021. Wymer 2021, VII reflects that the COVID-19 pandemic has accelerated the process of digitization of academic institutions' resources.

81 https://about.jstor.org/news/68000-additional-free-articles-added-to-early-journal-content.

82 https://about.jstor.org/covid19.

83 Galimberti 2009, 169. ["The hard sciences (with international dissemination) have long since turned to disciplinary archives (Pubmedcentral, Arxiv, etc.), the humanities (usually with national dissemination), on the other hand, are somewhat at a standstill"]

84 Mordenti 2012, 175 and 180.

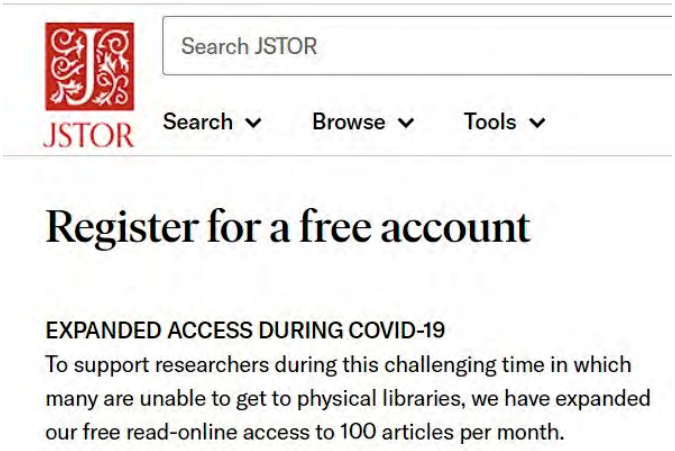


Fig. 6: The expanded access at *JSTOR*.

7 Concluding remarks. Bringing humanists back to the centre

> Dunque forse dobbiamo cominciare a pensare (per entrare davvero nella "seconda fase" della filologia informatica) a [...] un'edizione che ha più padri, o madri, e che non appartiene in esclusiva a nessuno, perché è frutto di una sorta di cervello collettivo, della comunità scientifica, della "comunità degli interpreti."[85]
>
> Co-creation is one of the founding features of the digital turn in the human sciences.[86]

Seneca said: *ducunt volentem fata, nolentem trahunt*,[87] the fates lead with them those who accept them and drag along the recalcitrant. Similarly, working for the consolidation of the *Open Digital Humanities* means trying not to passively undergo the telematics revolution, but to exploit its potential. The aim of an *open* approach would be the consolidation of interoperability, just as P. Monella suggested at the end of my presentation at the

85 Mordenti 2012, 182 ["So perhaps we need to start thinking (to really enter the "second phase" of computer philology) about [...] an edition that has multiple fathers, or mothers, and that does not belong exclusively to anyone, because it is the result of a kind of collective brain, of the scientific community, of the "community of interpreters""].

86 Drucker – Lunenfeld – Presner – Schnapp 2008.

87 Sen. *Epist.* 107, 11, 5.

conference that gave rise to this volume. Interoperability is related to the concept of co-creation mentioned by the UCLA team referring to DH and by R. Mordenti in the contest of Digital Philology.[88]

My proposal for the future is to bring Humanities back to the centre, just like a hinge between digital and traditional media, between copyright protection and *open access* to knowledge. An impetus in this direction seems to be the XXVIII Nestle-Aland edition of the Greek New Testament,[89] which supplements the printed edition, protected by *copyright-all right reserved*, with an interactive electronic version[90] in OA and with a virtual room[91] in Open Source where scans of ancient manuscripts stored all over the world can be consulted and their transcriptions and analyses are edited by the scientific community through a collaborative system based on the *wiki* model,[92] just as *Papyri.info* works.[93]

Paraphrasing the motto of John Lasseter, visionary founder of Pixar studios, we could say that *humanities challenge technology and technology inspires humanities*[94] to achieve what we have defined at the beginning of our journey as *cultural informatics*.

Bibliography

Angelopoulos, C. (2022), *Study on EU Copyright and Related Rights and Access to and Reuse of Scientific Publications, Including Open Access – Exceptions and Limitations, Rights Retention Strategies and the Secondary Publication Right*, Luxembourg, https://data.europa.eu/doi/10.2777/891665.

Bauman, Z. (2000), *Liquid modernity*, Cambridge (UK) – Malden.

Berners-Lee, T. (1990), *Information Management: A Proposal*, CERN, https://cds.cern.ch/record/369245/files/dd-89-001.pdf.

Bodard, G. – Garcés, J., (2009), *Open Source Critical Editions: A Rationale*, in *Text Editing, Print, and the Digital World*, ed. by M. Deegan – K. Sutherland, Aldershot, 83–98.

Braudel, F. (1958), *Histoire et Sciences sociales : La longue durée*, Annales 13, 725–3. [https://www.persee.fr/doc/ahess_0395-2649_1958_num_13_4_2781].

Brunner, T. F. (1991), *The* Thesaurus Linguae Graecae*: Classics and the Computer*, Library Hi Tech 9/1, 61–7.

Burdick, A. – Drucker, J. – Lunenfeld, P. – Presner, T. – Schnapp, J. (2012), *Digital Humanities*, Cambridge (MA).

Busa, R. (1980), *The Annals of Humanities Computing: The Index Thomisticus*, Computers and the Humanities 14, 83–90. [https://www.jstor.org/stable/30207304]

88 Cf. in this regard Bodard – Garcés 2009: "Open Source Critical Editions are more than merely presentations of finished work; they involve an essential distribution of the raw data, the scholarly tradition, the decision-making process, and the tools and applications that were used in reaching these conclusions".

89 Cf. Paulson 2021.

90 Digital Nestle-Aland (http://nestlealand.uni-muenster.de).

91 New Testament Virtual Manuscript Room (http://ntvmr.uni-muenster.de)

92 Cf. https://www.treccani.it/enciclopedia/wiki.

93 Cf. Reggiani 2017, 232–40.

94 Cf. Neupert 2016, 171: "The art challenges the technology, and for its part, technology inspires the art. It is a marvelous balancing act".

Busa, R. (1992), ed., *Thomae Aquinatis opera omnia: cum hypertextibus in CD-ROM*, Milan.
Busa, R. (2004), *Foreword: Perspectives on the Digital Humanities*, in Schreibman – Siemens – Unsworth 2004, XVI–XXI.
Ciotti, F. (2018), *Dall'Informatica umanistica alle Digital Humanities. Per una storia concettuale delle DH in Italia – DH2018*, in *2018 Digital Humanities Conference, Mexico City*, https://dh2018.adho.org/en/dallinformatica-umanistica-alle-digital-humanities-per-una-storia-concettuale-delle-dh-in-italia.
Ciotti, F. – Corradini, E. – Cugliana, E. – D'Agostino, G. – Ferroni, L. – Fischer, F. – Lana, M. – Monella, P. – Roeder, T. – Rosselli Del Turco, R. – Sahle, P. (2022), *Digital Scholarly Editions Manifesto*. Umanistica Digitale, 6, 103–8. [https://doi.org/10.6092/issn.2532-8816/14814]
Ciotti, F. – Crupi, G. (2012), curr., *Dall'Informatica umanistica alle culture digitali. Atti del convegno di studi (Roma, 27-28 ottobre 2011) in memoria di Giuseppe Gigliozzi*, Rome, https://www.editricesapienza.it/node/7688.
Crane, G. – Bamman, D. – Jones, A. (2013), *ePhilology: When the Books Talk to Their Readers*, in *A Companion to Digital Literary Studies*, ed. by S. Schreibman – R. Siemens, https://doi.org/10.1002/9781405177504.ch2.
D'Andrea, V. – Toccoli, S., (2010), *Dall'Open Source all'Open Access: ideologie, percorsi ed intersezioni*, in *Accesso aperto alla conoscenza scientifica e sistema trentino della ricerca: atti del Convegno tenuto presso la Facoltà di Giurisprudenza di Trento il 5 maggio 2009*, ed. by R. Caso – F. Puppo, Trento, 39–51. [http://eprints.biblio.unitn.it/1821/]
Drucker, J., (2021), *The Digital Humanities Coursebook: An Introduction to Digital Methods for Research and Scholarship*, Abingdon – New York.
Drucker, J. – Lunenfeld, P. – Presner, T. – Schnapp, J. (2008) *A Digital Humanities Manifesto*, web.archive.org/web/20090906003307/http://dev.cdh.ucla.edu/digitalhumanities.
Elkin-Koren, N. (2001), *Of Scientific Claims and Proprietary Rights: Lessons from the Dead Seas Scrolls Case*, Houston Law Review 38/2, https://houstonlawreview.org/article/4068-of-scientific-claims-and-proprietary-rights-lessons-from-the-dead-seas-scrolls-case.
Fiormonte, D. (2003), *Scrittura e filologia nell'era digitale*, Turin.
Galimberti, P. (2009), *Open Access: principali ostacoli per un'ampia diffusione in Italia*, Informatica e diritto 18, 2161–70, http://www.ittig.cnr.it/EditoriaServizi/AttivitaEditoriale/InformaticaEDiritto/Galimberti.Ied.2-2009.html.
González-Solar, L. – Fernández-Marcial, V. (2019), *Sci-Hub, a Challenge for Academic and Research Libraries*, El Profesional de la Información 28/1, https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/epi.2019.ene.12.
Hockey, S. (2004), *The History of Humanities Computing*, in Schreibman – Siemens – Unsworth 2004, 3–19.
Kuhn, T. S. (1970), *The Structure of Scientific Revolution*, Chicago – London.
Lana, M. (2012), *Da una* digital library *del latino tardo ad un* corpus *globale*, in Ciotti – Crupi 2012, 134–50.
Lessig, L. (2013a), *Aaron's Laws: Law and Justice in a Digital Age*, public lecture, Harvard Law School. Poster: https://ethics.harvard.edu/event/aarons-laws-law-and-justice-digital-age. Video: https://www.youtube.com/watch?v=9HAw1i4gOU4. Transcript: http://web.archive.org/web/20170718215941/http://www.correntewire.com/transcript_lawrence_lessig_on_aarons_laws_law_and_justice_in_a_digital_age.
Lessig, L. (2013b), *Why They Mattered: Aaron Swartz*, PoliticoMagazine, 22 December 2013, https://www.politico.com/magazine/story/2013/12/aaron-swartz-obituary-101418.
Lovink, G. (2008), *Zero Comments: Blogging and Critical Internet Culture*, New York.
Magnani, M. (2018), *The Other Side of the River: Digital Editions of Ancient Greek Texts Involving Papyrus Witnesses*, in *Digital Papyrology II*, ed. by N. Reggiani, Berlin – Boston, 87–102.
Margoni, T. – Perry, M. (2011), *Scientific and Critical Editions of Public Domain Works: An Example of European Copyright Law (Dis)Harmonization*, Canadian Intellectual Property Review 27/1, 157–70. [https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1961535]
McGann, J. J. (1992), *A Critique of Modern Textual Criticism*, Charlottesville.
McGann, J. J. (2002), *Literary Scholarship in the Digital Future*, in *The Chronicle of Higher Education*, https://www.chronicle.com/article/literary-scholarship-in-the-digital-future.
McLuhan, M. – Fiore, Q. – Agel, J. (1996), *The Medium is the Massage: An Inventory of Effects*, San Francisco.

Milanese, G. (2020), *Filologia, letteratura, computer: idee e strumenti per l'informatica umanistica*, Milan.
Mordenti, R. (2012), *Domande teoriche sul concetto di edizione (nel nome di Giuseppe Gigliozzi)*, in Ciotti – Crupi 2012, 167-82.
Neupert, R. (2016), *John Lasseter*, Urbana – Chicago – Springfield.
Nimmer, D. (2001), *Copyright in the Dead Sea Scrolls: Authorship and Originality*, Houston Law Review 38/1, https://houstonlawreview.org/article/4015-copyright-in-the-dead-sea-scrolls-authorship-and-originality.
Numerico, T. – Fiormonte, D. – Tomasi, F. (2010), *L'umanista digitale*, Bologna.
Oakes, J. L. (2001), *The Dead Sea Scrolls: A Live Copyright Controversy*, Houston Law Review 38/1, https://houstonlawreview.org/article/4016-the-dead-sea-scrolls-a-live-copyright-controversy.
Orlandi, T. (2012), *Per una storia dell'Informatica Umanistica*, in Ciotti – Crupi 2012, 49-102.
Orlandi, T. – Tomasi, F. (2023), *Una storia dell'informatica umanistica in Italia*, in *Digital humanities: metodi, strumenti, saperi*, ed. by F. Ciotti, Rome, 35–47.
Pantelia, M. (2000), *'Noῦς into Chaos'. The Creation of the Thesaurus of the Greek Language*, International Journal of Lexicography 13/1, 1–11. [https://doi.org/10.1093/ijl/13.1.1]
Pantelia, M. (2003), *The Thesaurus Linguae Graecae Project: Looking towards the 21st century*, in *Ἡ λεξικογραφία της αρχαίας, μεσαιωνικής και νέας ελληνικής γραμματείας = The Lexicography of Ancient, Medieval and Modern Greek Literature*, ed. by I. N. Kazazēs, Thessalonikē, https://www.greek-language.gr/greekLang/files/document/conference-1997/02-en-153-158.pdf.
Paulson, G. S. (2021), *The Nestle-Aland as Open Digital Edition: Already and Not Yet*, in *Ancient Manuscripts and Virtual Research Environments* (Classics@ 18), ed. by C. Clivaz – G. V. Allen, Washington, https://classics-at.chs.harvard.edu/classics18-paulson/.
Reeve, M. D. (1999), *Textual Criticism*, in *The Oxford Classical Dictionary*, ed. by S. Hornblower – A. Spawforth, Oxford, 1490–1.
Reggiani, N. (2017), *Digital Papyrology I*, Berlin – Boston.
Reggiani, N. (2021), *Papyrology in the Time of Coronavirus: Some Reflections on the Impact of COVID-19 Emergency on Papyrological Research*, unpublished paper originally presented at the Berliner Papyrologisches Kolloquium, 3 December 2020.
Sappa, C. – Widła, B. (2023), *Framing Texts and Images: Critical and Posthumous Editions in the Digital Single Market*, International Review of Intellectual Property and Competition Law 54, 1359–80. [https://doi.org/10.1007/s40319-023-01394-9]
Scholes, R. – Wulfman, C. (2008), *Humanities Computing and Digital Humanities*, South Atlantic Review 73/4, 50–66. [https://www.jstor.org/stable/27784811]
Schreibman, S. – Siemens, R. – Unsworth, J. (2004), eds., *A Companion to Digital Humanities*, Malden – Oxford – Carlton, https://onlinelibrary.wiley.com/doi/book/10.1002/9780470999875.
Stallman, R. (1985) *The GNU Manifesto*, http://www.gnu.org/gnu/manifesto.html.
Stallman, R. (2009), *Why "Open Source" Misses the Point of Free Software*, Communications of the ACM 52/6, 31–3. [https://dl.acm.org/doi/10.1145/1516046.1516058].
Tissoni, F. (2010), *L'editoria multimediale del nuovo Web. Semantic Web e Web 2.0*, Milan.
Tissoni, F. (2015), *Pour un corpus numérique comparatiste des traductions d'Homère*, Corpus Eve 2, http://journals.openedition.org/eve/1273.
Tuccillo, M. (2019) *Il ruolo del provider negli illeciti in materia di diritto d'autore alla luce della nuova direttiva sul copyright*, Rivista di Diritto delle Arti e dello Spettacolo 2, 107–26.
Varvaro, A. (1999a), *Il testo letterario*, in *Lo spazio letterario del Medioevo. Il Medioevo volgare*, *I (La produzione del testo)*, tomo 1, ed. by P. Boitani – M. Mancini – A. Varvaro, Rome, 387–422.
Varvaro, A. (1999b), *The "New Philology" from an Italian Perspective*, transl. by M. Cherchi, Text 12, 49–58. [https://www.jstor.org/stable/30228024]
Varvaro, A. (2012), *Prima lezione di filologia*, Bari.
Wiener, N. (1989), *The Human Use of Human Beings: Cybernetics and Society*, London.
Wymer, K. C. (2021), *Introduction to Digital Humanities: Enhancing Scholarship with the Use of Technology*, New York.

Fausto Pagnotta

The History of Political Thought and the AISPP Website in the 'Post-Truth' Era

Huizinga's Lesson and Some Insights from Digital Papyrology

1 Information criticism in historical research

The work of the historian, specifically the historian of political thought, has always involved the critical-analytical study of documentary sources, thanks to which it is possible to tentatively propose reconstructions and interpretations – always partial, it goes without saying – of the past and of the political concepts transmitted by it, starting from the earliest ones, analyzed and contextualized in their various historical, cultural, social, political, legal, economic, and scientific aspects. As Luciano Canfora has emphasized, this underlines the importance of the historian's ability/competence to "categorically distinguish on the ground of categories between the multiple types of documents, which are" never "absolutely susceptible to a single judgment of acquittal or condemnation [...] in recognizing the provisional nature of any historiographical reconstruction around most past events."[1] All this takes place within an epistemic perspective of research, which is necessarily aimed, on the level of methodological approach, at "bringing out the 'doubt' and the process of constant questioning that animates the practice of history."[2] This is done to avoid or at least contain overly simplistic, distorted, or ideologically biased interpretations of historical facts. Indeed, as Georges Duby aptly stated, "History gives 'lessons' to the extent that it teaches methodological doubt and rigor, as it is training in information criticism."[3]

This preparatory purpose, which is transversal across multiple scientific disciplines, has as its indispensable prerequisite particular attention and care, and therefore the safeguarding, of those "materials of time," as have been significantly defined, for example, "the book-men, the ethnographic objects, the statues, the square,"[4] to which

1 Canfora 2013, 58–61. ["distinguere categorialmente tra i molteplici tipi di documenti, i quali non sono» mai «assolutamente passibili di un unico giudizio d'insieme di assoluzione o di condanna (...) nella constatazione della provvisorietà di qualunque ricostruzione storiografica intorno alla gran parte degli eventi del passato"]. Hyperlinks last accessed on 13.7.2024.

2 Genovesi 2002, 41. ["emergere il 'dubbio' e il processo di costante interrogazione che anima fare storia"]

3 Duby 1986, 182. ["La storia dà 'insegnamenti' nella misura in cui insegna il dubbio metodico, il rigore, in quanto è addestramento a una critica dell'informazione"]

4 Papagno 2000, 27; more generally, see also pp. 24–9, 36–8. ["gli uomini-libro, gli oggetti etnografici, le statue, la piazza"]

https://doi.org/10.1515/9783111070162-005

we can certainly add textual sources. Such "materials of time" allow us to translate time itself, and thus temporality, into a process of signification that enables the birth of a history and, through the historian's research work, of *the* history, critically understood and examined, always open to possible interpretations and reinterpretations, and always exposed, as Reinhart Koselleck emphasized, to "two variants – the objective side and the subjective side – that logically exclude each other" and that "confer an ambivalence to the concept [*sc.* of history], which since then [*sc.* Since the meaning it assumed in the modern era starting from the end of the 18th century] has remained inherent in it and from which derive its applicability as a watchword, its predisposition to ideology," and at the same time "to the criticism of ideology."[5]

It is indeed to avoid turning into ideology that the historian's research work cannot but conceive historical research as never closed in its assumptions as well as its conclusions. This can begin with the ability to relate time itself, which denotes "a category with a general dimension," in a "clear and precise [manner] to the 'materials' in which it is 'visible' and 'treatable'."[6] It is precisely by passing through this 'materiality,' in which time, with its factual progression, can become history, in the interpretive reading given by whom does historical research, that it is created "a direct relationship between the temporal 'dimension,' the process of historical identity of a group, and the *materials* involved."[7] Among the "materials of time," the written texts, transmitted through the most diverse supports, remain among the privileged "materials" of the historian's research work, including that of the historian of political thought, with the awareness that "writing history is also, and not secondarily, giving life to a *narrative fabric* aimed at connecting, by giving them meaning, the factual segments that the available or known documentation offers," a work of "connection," therefore, which is and remains, in fact, "*intrinsically conjectural*."[8]

It is in this context of study and research that the reconstruction and critical examination of documentary sources, whether textual or otherwise, assume paramount importance – as does their accessibility to both experts and novices, because every "historical narrative is based on a stimulating and fruitful tension between the historian's

5 Koselleck 2009, 24. ["due varianti – il lato oggettivo e quello soggettivo – , che logicamente si escludono l'un l'altra" e che "conferiscono al concetto [*sc.* di storia] un'ambivalenza che da allora [*sc.* dal significato assunto in epoca moderna a partire dalla fine del XVIII secolo] rimane insita in esso e da cui derivano la sua applicabilità come parola d'ordine, la sua predisposizione all'ideologia", come al contempo "alla critica dell'ideologia"]

6 Papagno 2000, 27. ["una categoria a dimensione generale", in modo "ben chiaro e preciso ai 'materiali' in cui esso è 'visibile' e 'trattabile'"]

7 Papagno 2000, 27. ["una relazione diretta tra la 'dimensione' temporale, il processo d'identità storica d'un gruppo e i *materiali* implicati"]

8 Canfora 2013, 62. ["lo scrivere storia è anche, e non secondariamente, dar vita a un *tessuto narrativo* volto a connettere, dando loro un senso, i segmenti fattuali che la documentazione disponibile o conosciuta offre", un lavoro "di connessione", dunque, che di fatto è e rimane "*intrinsecamente congetturale*"]

interpretation, and the materials that fuel it and constitute its foundation, providing evidence for it."[9]

2 Information criticism and 'post-truth'

Taking up Georges Duby's words, it is precisely in the historian's research work – specifically, in that of the historian of political thought, who studies the semantic evolution of political concepts in history, their identification, reconstruction, and contextualization, both synchronically and diachronically – that the "training in information criticism"[10] proves to be central in terms of methodological correctness. Today, indeed, in a global society that is increasingly digitized worldwide due to the progressive and constant diffusion of the Internet, of the Information and Communication Technologies (ICTs), and of Artificial Intelligence (AI), historical information, in its textual documentary forms, which materially sediment within the digital space of the Web, can be potentially exposed to the constant risk of replication, interpolation, decontextualization, manipulation, counterfeiting, and falsification. Indeed, as has been appropriately noted, in the Web "the hierarchies of relevance in the relationship between narrative and sources, on which the forms of communication of historical discourse have traditionally been built, seem [...] to be called into question and, more or less profoundly, reconfigured."[11]

Adding to this is the fact that every digital content, in its production, publication, and dissemination on the Web through widely available and accessible digital sharing platforms such as the Social Network Sites, is firstly characterized, as Dana Boyd highlighted,[12] by five specific features: (1) "persistence" (every piece of content spread online can be permanently recorded and archived); (2) "replicability" (any content spread online can potentially be copied and transferred from one medium to another, as well as from one context to another, increasing its persistence, shareability, but also the risk of disseminating content altered from its original version, even without the original author's intentional modification); (3) "searchability" (any informational content spread online can be traced through specific search tools); (4) "scalability" (the Web redefines the scale of dissemination of any content published on it, which can thus achieve a high potential for visibility and amplification from micro to macro scale, at a global level);

9 Vitali 2004, 120. ["...racconto storico si fonda su una stimolante e proficua tensione fra l'interpretazione dello storico e i materiali che l'alimentano ne costituiscono il presupposto, ne forniscono le prove"]

10 Duby 1986, 182. ["addestramento a una critica dell'informazione"]

11 Vitali 2004, 120. ["le gerarchie di rilevanza nel rapporto fra narrazione e fonti, sulle quali sono state tradizionalmente costruite le forme di comunicazione del discorso storico, sembrano (...) essere rimesse in discussione e, più o meno profondamente, riconfigurate"]

12 See Boyd 2009.

(5) "delocalizability" (on one hand, any Web user can be delocalized from a specific physical point/place in space – e.g., home, university, research lab – into the digital environment through digital devices; on the other hand, through digital location-tracking technologies, these 'located' points/places assume particular relevance[13]).

In such a context, which is communicative, informational, socio-relational, and cultural at the same time, the digital space tends to create conditions in which potentially every piece of content produced, expressed, or disseminated within it may undergo distortions and decontextualizations with respect to the its original author's meaning and intentions and to the initial context in which it was created. This is a risk factor that can affect any information disseminated on the Internet as well as any document present in the digital space. For example, the authorial texts, particularly those in the history of political thought, can be copied, altered, interpolated, manipulated, and falsified, with millions of Internet users, especially those not well-versed in the discipline, often lacking the cultural and critical tools necessary to recognize these manipulative interventions on authoritative texts.

This phenomenon is referred to as "information disorder,"[14] which spreads across the Web mainly through three concepts/behaviors: (1) disinformation (one or more users/issuers intentionally creating and spreading false or distorted information online to confuse the public or damage the reputation of specific social, poli-tical, economic, scientific, cultural, ethnic, and religious targets); (2) misinformation (the creation or dissemination of false or incorrect information online, often amplified by social media, without a specific intent to harm[15]; (3) malinformation (the illicit dissemination of truthful but confidential information, including phenomena like hate speech and mudslinging, which manifest in both private and public-poli-tical spheres).[16]

These three concepts/behaviors – disinformation, misinformation, and malinformation – tend to encompass the concept of 'fake news'[17] in its various forms and contribute to the formation of the semantic domain of the concept of 'Post-Truth',[18] which has acquired a polysemous nature that primarily indicates the idea of a context where the concept of 'truth' is considered unimportant, if not irrelevant. As Michele Sorice has

13 This is a paradox, since it means that every user connected to the Internet results to be more or less connected to the physical location they are in. Applied to content disseminated on the Web, this makes it potentially visible, usable, and relevant regardless of the physical location of its production and online posting.

14 For a clear synthesis see Sorice 2022, 172–81; for further observations see e.g. McIntyre 2019; see also Nicita 2021; Quattrociocchi – Vicini 2016.

15 This phenomenon is facilitated by the rapid spread of digital content on the Web, leading to transcription errors and deficiencies in accurate source control.

16 See e.g. Sorice 2011; Warlde – Derakhshan 2017.

17 See Sorice 2022, 177; also Riva 2018; Orecchia – Preatoni 2022.

18 For a deeper discussion of the topic, see McIntyre 2019 with relevant bibliography; see also Ferraris 2017; Lorusso 2018; Quattrociocchi – Vicini 2018.

well summarized, this reflects "a social trend in which objective and/or verifiable facts are less significant and important than appeals to the emotional sphere, pre-existing personal beliefs, and, in general, unverified social and media 'narratives'."[19] Moreover, between 2016 and 2018, the concept of 'post-truth' came to signify the impossibility of "discernment," as "the speed of the media (which has in fact become 'instantaneity') would have made it impossible to distinguish the true from the false."[20]

We are indeed in a context – the 'post-truth' era – that is characterized by potential progressive informational pollution and a weakening of trust in a scientifically grounded critical spirit. Given this context, before proceeding in our discourse, it may be useful to revisit the insightful analysis developed in 1935 by the great Dutch historian Johan Huizinga,[21] which is, in many ways, paradigmatic and prophetic with respect to certain mass behavioral dynamics that characterize the digital 'post-truth' era we are currently experiencing.

3 Huizinga's lesson

In his work titled *In de schaduwen van morgen. Een diagnose van het geestelijk lijden van onzen tijd* ("In the Shadow of Tomorrow: A Diagnosis of the Spiritual Suffering of Our Time", translated into Italian by Einaudi in 1937 as *La crisi della civiltà* ["The Crisis of Civilization"]), Huizinga, in chapters seven and eight, dealing with "The general weakening of reason" and "The decline of critical spirit," denounced the infiltration of a "general weakening of reason" in Western civilization and culture. This was due to what he called "visual suggestibility" – the susceptibility to images – through which, for example, "advertising," he said, "grabs the modern man and strikes him at the weak point of his diminished capacity to judge."[22] This process could be triggered by mass media "for both commercial advertising and political propaganda," targeting – according to the Dutch historian – the most vulnerable aspect of modern man: the emotional side, often awakening "the thought of satisfying a desire."[23]

19 Sorice 2022, 175. ["una tendenza sociale in cui i fatti oggettivi e/o verificabili risultano meno significativi e importanti dei richiami alla sfera emozionale, alle convinzioni personali pregresse e, in generale, a 'narrazioni' sociali e mediatiche non verificate"]

20 Sorice 2022, 175. ["la velocità dei media (diventata di fatto 'istantaneità') avrebbe reso impossibile distinguere il vero dal falso"]

21 It is worth noting that Johan Huizinga (Groningen 1872 – De Steeg, Arnhem, 1945), as early as 1933, foresaw the consequences that National Socialism could bring to Europe and to the very idea of civilization. Due to his opposition to National Socialism, Huizinga was arrested in 1942 and, held prisoner in De Steeg, a place near Arnhem, he died there on February 1, 1945.

22 Huizinga 2012, 45. ["la pubblicità afferra l'uomo moderno, e lo colpisce nel lato debole della sua diminuita capacità di giudicare"]

23 Huizinga 2012, 45. ["per la pubblicità commerciale come per la propaganda politica" … "il pensiero della soddisfazione di un desiderio"]

Indeed, the power of a mental image, created by a sequence of images or words, or both together, characterized by a high "sentimental" (emotional) component, establishes – according to Huizinga – in those exposed to it a "state of mind" that does not remain as such but leads to the "formation of a judgment, which is made [...] in a rapid instant." This speed, combined with the emotional conditioning underlying it, precludes careful and accurate critical scrutiny by rational thought, which conversely requires time to develop. The Dutch historian noted that all of this occurs within a mass media context where "notions of all kinds, to an extent never thought of before, and arranged in ways never imagined before, are made available to the masses."[24] He in fact registered the advent of the mass media society, a precursor to what – with the spread of the Internet and the Web – Manuel Castells defined in the mid-1990s as the Network Society,[25] and today, with the global diffusion of Social Network Sites (SNS), the Social Network Society.[26]

The exponential increase in the amount of information to which individuals are continually subjected and stimulated, along with the speed at which this information reaches the recipients, and the predominance of the emotional response over the logical-rational one – these are all characteristics that Huizinga noted in 1935 and that we can well observe today in the Social Network Society and the so-called digital 'post-truth' era. Huizinga believed that this constant exposure to exorbitant, often chaotic, flows of information, facilitated by the progressive blending of text and images, without the possibility of critical scrutiny due to the rapid "cognitive bombing and overload" (now defined as "information overload"[27]) with which they reach the "media consumers," contributed to the emergence of a "weakening of critical passion, a muddling of critical power," and ultimately "a decline in the need for truth." This phenomenon, he observed, affected not only the "consumers of doctrine" but also began to touch the "producers of doctrine,"[28] the men of science. He noted that in his time "the need to think exactly and objectively, as much as possible, about things graspable by reason, and to critically examine the thought itself, is becoming weaker," and without ever succumbing to rationalist temptation, he observed that in such a mass media context, "every boundary between logical, aesthetic, and affective functions" was "intentionally neglected," leading to the frequent confusion of "the suggestions of interest and desire with conviction based on knowledge."[29] Huizinga concluded by denouncing that in his time, the "renun-

24 Huizinga 2012, 46. ["formazione di un giudizio, che si compie (...) in un rapido istante" ... "Nozioni d'ogni genere, in una misura mai pensata finora, e allestite in modi non mai immaginati, vengono messe a portata delle masse"]

25 See Castells 1996; see also Castells 1997, 2001, 2009.

26 See Boccia Artieri 2012.

27 See e.g. Strother – Fazal – Ulijn 2012.

28 Huizinga 2012, 48.

29 Huizinga 2012, 50. ["Il bisogno di pensare esattamente e obiettivamente, quanto si può, intorno alle cose afferrabili dalla ragione, e di vagliare criticamente il pensiero stesso, si fa più debole" ... "Ogni delimitazione

ciation of the veto of criticism" by rational thought could be "illustrated most effectively" by the emergence and spread, also and especially through the mass media, of pseudoscientific theories like the "theory of race," the basis of what he called the "racist doctrine,"[30] scientifically baseless but serving as a "self-apology" with the intent of "elevating" some people "above others, and at the expense of others," being inherently "always belligerent, hostile, anti-something (anti-Asian, anti-African, anti-proletarian, or anti-Semitic)."[31] This "racist theory," which, as Huizinga wrote, managed to garner support from part of the scientific community and of the people where it was propagated, provided the theoretical basis for the ideological-propagandistic legitimization of racist and supremacist demands that led to the Holocaust.

I believe that Huizinga's analysis on the potential consequences of the massive and chaotic spread of pseudoscientific information in a mass media society is particularly relevant in the current so-called 'post-truth' era. Despite the undeniable advantages brought by the Internet, social media, and ICTs in various aspects (social-relational, communicative, scientific-cultural, economic-labor), these platforms often serve as incubators and disseminators of manipulated, false, pseudoscientific, or even non-scientific information on a global scale, largely unchecked except for fact-checking or debunking platforms.

4 'Post-truth' and disinformation

In this context, it seems essential to linger exploring the meaning of the neologism 'post-truth,' which has entered the Web and all media in general, so much so that the Oxford Dictionaries selected it as the word of the year in 2016, particularly after "Brexit" and the election of Donald Trump as President of the United States. According to the Oxford Dictionaries, "*Post-truth* is an adjective defined as 'relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief'."[32] Moreover, although recording the frequent attestations of the term 'post-truth' in its literal meaning of "after the truth is known," the Oxford Dictionaries indicate that the first occurrence of 'post-truth' in the sense of "truth that has become irrelevant" was in 1992, when Serbian-American playwright Steve Tesich, in an article for "The Nation" journal, wrote about the "Iran-Contra Af-

di confine tra le funzioni logiche, quelle estetiche e quelle affettive» fosse «intenzionalmente negletta" … "le suggestioni dell'interesse e del desiderio con la convinzione fondata su una conoscenza"]

30 Huizinga 2012, 51–2. ["dottrina razzistica"]

31 Huizinga 2012, 53. ["sempre bellicosa, ostile, anti-qualcosa (antiasiatica o antiafricana o antiproletaria o antisemita)"]

32 Oxford Languages 2016.

fair"[33] and the Gulf War: "we, as a free people, have freely decided that we want to live in some post-truth world."[34]

In Italy, as noted by the Accademia della Crusca in an insightful article by Marco Biffi,

> one of the earliest attestations of *post-verità* (the first found so far) is in an article published in "La Repubblica" on May 1, 2013, written by Barbara Spinelli, about the Gulf War: "It will be subversive truth, Letta says, but on the contrary we are still immersed in what has been called – since Bush has started war in Iraq – the post-truth era: of euphemisms that make facts beautiful, of words contrary to what they mean." Here, we see early sectorial uses; by 2016, the word had become virally common. In Italian, *post-verità* has been used both as an adjective and as a noun from its earliest attestations, due to the peculiar transformations functional for adaptation. The English phrases where it is more frequently found (*post-truth politics*, *post-truth society*, *post-truth era*) favour the transition to a noun for the Italian morphological rules. For example, the aforementioned *post-truth world* naturally becomes "mondo della *post-verità*" rather than "mondo *post-verità*." Similarly, Spinelli's *era della post-verità* reflects a *post-truth era*, with *post-truth* as an adjective. The use of *post-verità* as a noun has been contested by some (based on the specific meaning of *post-truth* in English), but it is now widespread on the Web and in newspapers, with reference to a pseudo-truth based on emotions and personal beliefs at the expense of objective facts. Rather, it seems to have become predominant, and in almost all contexts and meanings where truth would be used, *post-verità* is employed (*la post-verità*, *le post-verità*, etc.), as in this text.[35]

The phenomenon of manipulation and falsification of information, documents, and textual sources, which today on the Web contributes to characterizing some of the main features of what has been defined as the "post-truth era," has always existed since ancient times.[36] However, in the digital space, just due to the aforementioned features that

33 Aa.Vv. 2024.

34 Oxford Languages 2016.

35 Biffi 2017, 73–4. ["una delle prime attestazioni di *post-verità* (la prima finora rintracciata) sia in un articolo apparso sulla 'Repubblica' il 1° maggio 2013, firmato da Barbara Spinelli, proprio in riferimento alla guerra del Golfo: 'Sarà verità sovversiva, dice Letta, e invece siamo tuttora immersi in quella che è stata chiamata – da quando Bush iniziò la guerra in Iraq – l'era della post-verità: degli eufemismi che imbelliscono i fatti, dei vocaboli contrari a quel che intendono'. Qui siamo di fronte a usi ancora settoriali; nel 2016 la parola è diventata viralmente comune. In italiano *post-verità* è usato fin dalle prime attestazioni sia con valore di aggettivo sia come sostantivo, proprio per le peculiari trasformazioni funzionali all'adattamento: i sintagmi inglesi in cui si ritrova più facilmente (*post-truth politics*, *post-truth society*, *post-truth era*) favoriscono infatti, per le regole morfologiche italiane, il trapasso al sostantivo. Si veda ad esempio il sopracitato *post-truth world*, che diventa più naturalmente 'mondo della post-verità' che 'mondo post-verità' (in cui sarebbe privilegiato il costrutto anglizzante, per altro in grande ascesa nella nostra lingua recente); e, d'altro canto, l'*era della post-verità* della Spinelli cela un *post-truth era*, con *post-truth* aggettivo. L'uso di *post-verità* come sostantivo è stato contrastato da alcuni (sulla base dello specifico significato che *post-truth* assume in inglese), ma è ormai molto diffuso sul web e sui giornali in riferimento alla pseudo-verità basata sull'emotività e sulle convinzioni personali a discapito dei fatti oggettivi; anzi, sembra ormai addirittura prevalente e con questo specifico significato è usato in quasi tutti i contesti e le accezioni in cui si potrebbe ricorrere a verità (*la post-verità*, *le post-verità*, ecc.), come del resto si è fatto anche in questo testo"]

36 See e.g. for the Roman world Segenni 2020.

each content produced or uploaded to the Web assumes, this phenomenon can have an unprecedented worldwide 'disinformative' impact over the millions of users of that informational dimension produced by the Web, digital technologies, and the Internet users themselves, a dimension significantly defined as the "infosphere."[37] The disinformative impact that can occur in cyberspace on any topic and through any content produced and uploaded to the Web can manifest through the so-called phenomenon of cybercascades. As Cass R. Sunstein has noted, these social and informational cascades "become more probable when information, even false," like the content that conveys it, "can reach hundreds, thousands, or even millions of people simply by pressing a button."[38] These informational cybercascades "influence our culture and even our way of thinking," often originating "within isolated communities that develop a particular inclination for certain products, movies, books, or ideas,"[39] with the aim of expanding their conceptions on the Web, which often lack recognition from the scientific community. In many cases, we witness on the Web the proliferation of fake scientific news – indeed we speak of the "society of pseudoscience" –, which can be observed in various fields of knowledge. Around such false information, proper communities of Internet users form, who, recognizing themselves as 'followers,' true 'believers,' of these beliefs, not accepted by the scientific community but raised to the status of absolute truths, develop specific identity profiles that often display a marked ideological character, not always conceptually organized but marked by some of the distinctive traits of ideology.

As Carlo Galli has significantly pointed out, indeed, ideology "has in itself a project of new objectivity, new humanity, and new order," and above all "it is committed with its supporters, the militants, [...] to remove the veil that distorts and prevents clear vision and to target those responsible for the obscurity." In this endeavor, ideology "is intrinsically polemic and almost always rejects for itself the very term 'ideology' and demands the qualification of 'doctrine,' or 'science,' obviously objective," constantly driven by "a polemic will for truth and struggle against error."[40] These are all characteristics – with different nuances depending on the context – that are found in those groups on the Web that organize around pseudoscientific conceptions and give rise to the phenomenon of the so-called 'echo chambers,' identified since the 1960s.[41] Today, in

37 See Floridi 2020.

38 Sunstein 2017, 127. ["diventano più probabili quando le informazioni, anche false, possono raggiungere centinaia, migliaia o addirittura milioni di persone premendo semplicemente un tasto"]

39 Sunstein 2017, 127. ["influenzano la nostra cultura e perfino il nostro modo di pensare (...) all'interno di comunità isolate, che sviluppano una particolare inclinazione per certi prodotti, film, libri o idee"]

40 Galli 2022, 8–14. ["ha in sé un progetto di nuova oggettività, di nuova umanità e di nuovo ordine" (...) è impegnata con i suoi fautori, i militanti (...) a togliere il velo che falsa e impedisce la visuale, e a colpire i responsabili dell'oscurità (...) è intrinsecamente polemica e rifiuta quasi sempre, per sé, il vocabolo stesso di 'ideologia' ed esige la qualifica di 'dottrina', o di 'scienza', ovviamente oggettiva (...) da una polemica volontà di verità e di lotta contro l'errore"]

41 See Key 1966.

relation to the Web, echo chambers consist of digital 'places' (public pages, private groups, shared discussions on major Social Network Sites, but also 'counter-information' forums, blogs created on spe-cific topics, sites often publishing conspiracy theories) "where communication relationships are characterized by a high degree of *homophily* concerning shared opinions, beliefs, facts, interpretations, and worldviews" through "semantically homogeneous information" that contributes "to the entrenchment of each individual in the seemingly dominant position,"[42] leading to the phenomenon of progressive "'group polarization,' that is, the phenomenon where prolonged discussion of the same thesis by those who declare themselves followers strengthens collective convictions, orienting them towards the most extreme and intransigent position."[43]

This attitude of radicalizing opinions/beliefs formed on the Web implies the affirmation of hyper-identity forms that prevent those who promote and experience them from conceiving the encounter with the other as a moment of enrichment and dialectical confrontation, under the illusion of remaining true to oneself and one's opinions/beliefs, while, as Jean-Pierre Vernant has stated, "to be oneself, it is necessary to project oneself towards what is foreign, to extend oneself into it and through it," because "remaining closed in one's own identity" – or what one believes to be such – "means losing oneself and ceasing to exist," as "one knows and constructs oneself through contact and exchange with the other."[44]

As I have explored elsewhere, in the Web, which "can be fully considered as a social environment, thanks to social networks that enhance its relational dimension,[45] […] we cannot avoid the task of continuing to recognize ourselves as relational beings, under penalty of progressive 'dehumanization' of the human being with respect to the representation […] he gives of himself and his fellow humans online."[46] Conversely, the phenomenon of online disinformation, supported by forms of polarization, radicalization, and identity closure, calls into question pre-cisely the socio-relational potential for confrontation and dialogue of the human.

---

42 Tipaldo 2019, 22–3, who on the topic of homophily in the Social Networks recalls McPherson – Smith-Lovin – Cook 2001; on the topic of echo chambers on the Web and their impact on information and online communicative relations see Del Vicario – Vivaldo – Bessi *et al.* 2016; Quattrociocchi – Vicini 2023.

43 Ainis 2018, 78. ["'polarizzazione di gruppo', ossia quel fenomeno per cui la discussione prolungata della stessa tesi, da parte di quanti se ne dichiarano seguaci, ha l'effetto di rafforzare i convincimenti collettivi, orientandoli verso la posizione più estrema e intransigente"]

44 Vernant 2005, 170. ["per essere se stessi, è necessario proiettarsi verso ciò che è estraneo, prolungarsi in esso e per mezzo di esso", perché "rimanere chiusi nella propria identità equivale a perdersi e a cessare di esistere», in quanto "ci si conosce e ci si costruisce mediante il contatto e lo scambio con l'altro"]

45 See Riva 2010.

46 Pagnotta 2018, 25. ["…può considerarsi a pieno titolo, grazie ai social network che ne implementano la dimensione relazionale, come un ambiente sociale (…) non possiamo esimerci dal compito di continuare a riconoscerci come esseri relazionali, pena una progressiva 'dis-umanizzazione' dell'essere umano rispetto alla rappresentazione (…) che in Rete dà di se stesso e degli altri suoi simili"]

No one can feel immune to online disinformation today, nor can any field of knowledge. Therefore, no scientific discipline can avoid taking specific responsibility for creating safe as well as open online spaces, where to make the results of their scientific research accessible, and to provide web users with documentary materials and sources that characterize their scientific production as well as cultural background. This also involves promoting activities not only for experts but also for all those potentially interested in the themes and scientific content of the same discipline. In light of these objectives, it is crucial that these digital 'places' for accessing scientific knowledge aim to both preserve and disseminate open-access textual sources and documentary materials related to specific disciplinary fields, ensuring they are secure and reliable from the perspective of critical-textual scientific analysis. Furthermore, they should strive to become reference points and community-sharing spaces in the digital realm for both experts and novices wishing to engage with the topics of each scientific discipline.

5 Some insights from digital papyrology

It is particularly in the goal of utilizing digital tools effectively for the analysis, preservation, and tradition of 'textual materials' with historical-documentary value that papyrology, since the mid-1960s, has progressively become a pioneer among the humanities in using computerized tools in its research.[47] This discipline indeed, as has been aptly highlighted, has always been based on instances of comparison and discussion on the methodological ground, and on quantitative and qualitative analysis to address the organizational and interpretive complexity of the fragments of 'materials of time,' such as papyri, which are constantly increasing in number and often dispersed across geographically distant collections, with the aim of reconstructing the documents they represent.[48] Papyrology is also engaged in the tasks of comparison and discussion of texts, with the goal of transcribing and interpreting their content from both paleographic and philological/textual perspectives, through a hermeneutic process in which each papyrologist must examine various and complex data to contextualize the objects of study.[49] With these disciplinary goals in mind, the computer and all the devices that digital technologies offer today are among the privileged tools for the daily work of the papyrologist. Thanks to these tools, papyrologists "can easily navigate the huge number of comparisons and bibliography, with which the study of even a single papyrus inevitably requires engagement."[50] For all these reasons, among the various humanities disci-

47 See Reggiani 2019a, 455–8.

48 See Reggiani 2019b, 11.

49 See Reggiani 2019b, 11.

50 Capasso 2005, 227. ["può orientarsi con una certa facilità nell'immensa mole di confronti e di bibliografia con cui lo studio di anche un singolo papiro lo porta a doversi più o meno inevitabilmente misurare"]

plines, papyrology finds indispensable work tools in digital resources for accessing primary texts, metadata, and images of papyri and related material.

Papryology is particularly relevant as a term of comparison within the current discourse because it has developed a platform for the online publication of its main data – the texts of the papyri and their related metadata, i.e., all contextual information – that is precisely aimed at achieving the right balance between the demands for open access, research sharing, and strict scientific control. The Papyri.info database indeed offers, in addition to the complete versions of the texts in the most recent editorial format – published under the Creative Commons Attribution 3.0 License (CC BY 3.0)[51] –, the possibility for any papyrologist to add or modify the textual editions under the careful supervision of an editorial board – a mechanism of updating and control that is always visible in the "Editorial History" and "All History" sections available on each record page. In this way, Papyri.info manages to offer the scientific community and any interested user the most updated and accurate version of the papyrological sources.[52]

Moreover, the community aspect is equally important for papyrology as for other disciplines. The Web and digital resources can make the scientific community more proximate by breaking down spatial and temporal barriers. To this end, promoting discussion and comparison among members of the scientific community is fundamental, thus papyrology leverages digital tools to manage and circulate relevant data and ensure their availability in terms of both sharing and accessibility.[53] The analogy between the digital resources and the human concept of *amicitia papyrologorum* – an established motto expressing traditional friendly scholarly cooperation and sharing – has repeatedly been stressed as a necessary and inherent component of the discipline.[54] Moreover, several online digital strategies – includ-ing institutional websites, a mailing list, and many personal blogs – have been developed to disseminate the papyrological scholarly research both within the specialists' community and the wider public.[55] Crowdsourcing projects have also been launched in order to involve the laypeople in the papyrological research, under the experts' supervision.[56]

Therefore, papyrology, with its prerogative of addressing complex interpretive contexts – both in the physical reconstruction of textual materials and their preservation,

51 https://creativecommons.org/licenses/by/3.0. This license allows to freely share ("copy and redistribute the material in any medium or format for any purpose, even commercially") and adapt ("remix, transform, and build upon the material for any purpose, even commercially") the provided material.

52 See Reggiani 2019b, 297 and 300–16.

53 See Reggiani 2019b, 12.

54 See Bagnall – Gagos 2007, 65; van Minnen 2009, 658; Sosin 2010; Bagnall 2012; Depauw – Broux 2016, 202, 210; Reggiani 2019b, 11–12, 20, 165, 262, 324, 337, 340, 358.

55 See Reggiani 2019b, 24–46.

56 In particular, consider the cases of the Ancient Lives project, temporarily suspended, aimed at deciphering papyri with palaeographical comparisons (see Reggiani 2019b, 30–1 and 223–6), and of the recent Vesuvius challenge, aimed at deciphering Herculaneum papyri with the help of Artificial Intelligence algorithms (see Reggiani 2024, 131–2).

deciphering, and interpretation, as well as in the orderly organization of increasing amounts of data and information made accessible through digital technologies – can offer valuable methodological insights applicable across various humanities disciplines, including the history of political thought, to foster a productive relationship between these disciplines and digital technologies.

6 The AISPP website

It is precisely in this perspective of enhancing the use of ICTs within specific scientific-disciplinary contexts that the new website (Fig. 1) of the Italian Association for the History of Political Thought (Associazione Italiana di Storia del Pensiero Politico, AISPP)[57] was inaugurated in 2023 by its President, Professor Francesco Tuccari (University of Turin), in collaboration with Vice-President Professor Francesca Russo (University of Naples "Suor Orsola Benincasa") and the AISPP Executive Council.

The project behind the creation of the new website for the AISPP aims to achieve several objectives, directed both internally towards its members and externally towards all those interested in the study of the history of political thought. Accordingly, the website's homepage features an "Activities" section, divided into subsections such as "Annual Conference," "Permanent Seminar," "Summer School," "Book of the Month," and "Call for Papers/Proposals," informing visitors about the activities organized and conducted by the AISPP, both in-person and online. To this end, the AISPP website provides information on events reported by members and published chronologically in the "Events" section within the main "Announcements" area (divided into "Events," "Publications," "Calls"), and it especially makes recorded event videos accessible through a dedicated "Media Library" section, accessible from the homepage. This section stores links to videos uploaded to the AISPP YouTube channel,[58] such as the book presentations featured in the "Book of the Month" section.

Furthermore, the "Publications" section within the "Announcements" area proves particularly useful in keeping updated on the scientific publications of AISPP members. Equally important, in the perspective of enhancing the network of resources provided online by the scientific community, is the "Journals" area, also accessible from the homepage. This section allows access not only to the Italian National Agency for the Evaluation of Universities and Research Institutes (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca, ANVUR) website but also to major scientific journals in the disciplinary field of the history of political thought or related sectors.

57 https://aispp.it.

58 https://www.youtube.com/@AISPP2023.

Fig. 1: AISPP website homepage.

The AISPP website aims therefore to serve as a reliable and accessible reference point for the entire scientific community, starting from those dedicated to political thought history studies, but its mission is also to become a solid scientific and cultural reference for all those interested in the discipline's themes, with a particular focus on the educational sector. To achieve this, the AISPP has chosen a clear and flexible graphical interface for its website, allowing for the future addition of new sections such as bibliographical repertories of relevant texts, as well as to make various scientific publications open-access, such as the proceedings of the AISPP National Conferences and publications from members who wish to share their work, along with critical editions of works by classic authors in the history of political thought.

In pursuing these goals, the AISPP, through its new website, aims for a "critical use of the Web," responding with scientific commitment to the pressing "needs of caution, methodological correctness, rigor in research and analysis of data and documents made available," particularly today, in the 'post-truth' era. It also underscores the urgent need – as paralleled by digital papyrology – "to take advantage in a serious and informed manner of what is newly available with respect to the first Web era,"[59] thanks to the

59 Minuti 2017, 17. ["esigenze di cautela, di correttezza di metodo, di rigore nella ricerca e nell'analisi dei dati e dei documenti resi disponibili (...) di avvalersi in modo serio e consapevole di quanto di nuovo, rispetto alla prima età del web, si è reso disponibile"]

development in the digital technologies and to what will be available in the near future thanks to Artificial Intelligence. There is an increasing integration of scientific knowledge, culture, and technology, which will lead to good results if the central importance of human relationships is maintained and emphasized.

7 Conclusions

Digital online platforms devoted to various scientific disciplines, such as the AISPP website for historians of political thought, should not only serve instrumental and informational purposes but also foster socio-relational aims in their being cultural and scientific environments intended to make research materials accessible and secure, and especially to create conditions for sharing the whole of those goods upon which the survival of every social and community context depends, as the scientific community itself should be regarded: the relational goods.[60] In this current digital society of 'post-truth,' where "any falsehood – especially thanks to the virality granted by social media – can propagate" globally "and potentially acquire the status of plausible truth,"[61] 'networking' by the scientific community – in our specific case, the community of the historians of political thought, but also, for the sake of comparisons, the community of the papyrologists – is crucial as it serves as a countermeasure against historical revisionism or negationism, as well as text and document manipulation or falsification, "activating trust and reciprocity spirit, that is social capital,"[62] thus promoting – both within and beyond the scientific community itself, as well as in the digital space – opportunities for debate, discussion, and accurate information, which are the only antidotes against any possible dogmatism or falsification in scientific and cultural matters.

Bibliography

Aa.Vv. (2024), *Iran-Contra Affair*, Written and fact-checked by the Editors of Encyclopedia Britannica, Encyclopedia Britannica, 18 Jun., https://www.britannica.com/event/Iran-Contra-Affair.

Ainis, M. (2018), *Il regno dell'Uroboro. Benvenuti nell'epoca della solitudine di massa*, Milan.

Bagnall, R. S. (2012), *The* Amicitia Papyrologorum *in a Globalized World of Learning*, in *Actes du 26e Congrès international de papyrologie. Genève 16–21 août 2010*, ed. by P. Schubert, Geneva, 1–5.

60 See Donati – Solci 2011; Donati 2019.

61 Serughetti 2019, XV. ["qualunque falsità – grazie soprattutto alla viralità garantita dai *social media* – ha la possibilità di propagarsi fino ad acquisire lo statuto di verità possibile"]

62 Folgheraiter 2015, 11. ["attivando fiducia e spirito di reciprocità, ossia capitale sociale"]

Bagnall, R. S. – Gagos, T. (2007), *The Advanced Papyrological Information System: Past, Present, and Future*, in *Proceedings of the 24th International Congress of Papyrology, Helsinki, 1–7 August, 2004*, ed. by J. Frösén – T. Purola – E. Salmenkivi, Helsinki, I, 59–74.

Biffi, M. (2017), *Viviamo dell'epoca della post-verità?*, Italiano Digitale 2 (luglio-settembre), 72–5, https://accademiadellacrusca.it/sites/www.accademiadellacrusca.it/files/page/2018/01/02/italiano_digitale_02.pdf.

Boccia Artieri, G. (2012), *Stati di connessione. Pubblici, cittadini e consumatori nella (Social) Network Society*, Milan.

Boyd, D. (2009), *Social Media is Here to Stay... Now What*?, Microsoft Research Tech Fest, Redmond, Washington, February 26, http://www.danah.org/papers/talks/MSRTechFest2009.html.

Canfora, L. (2013), *Noi e gli antichi. Perché lo studio dei Greci e dei Romani giova all'intelligenza dei moderni*, 4th ed., Milan.

Capasso, M. (2005), *Introduzione alla papirologia*, Bologna.

Castells, M. (1996), *The Rise of the Network Society*, in Id., *The Information Age: Economy, Society and Culture*, vol. I, Malden (MA, USA)-Oxford.

Castells, M. (1997), *The Power of Identity*, in Id., *The Information Age*: *Economy*, *Society and Culture*, vol. II, Malden (MA, USA)-Oxford.

Castells, M. (2001), *The Internet Galaxy: Reflections on the Internet, Business, and Society*, Oxford.

Castells, M. (2009), *Communication Power*, Oxford.

Del Vicario, M. – Vivaldo, G. – Bessi, A. – Zollo, F. – Scala, A. – Caldarelli, G. – Quattrociocchi, W. (2016), *Echo Chambers: Emotional Contagion and Group Polarization on Facebook*, Scientific Reports 6, https://doi.org/10.1038/srep37825.

Donati, P. (2019), *Scoprire i beni relazionali. Per generare una nuova socialità*, Soveria Mannelli (CZ).

Donati, P. – Solci, R. (2011), *I beni relazionali. Che cosa sono e cosa producono*, Turin.

Duby, G. (1986), *Il sogno della storia*, Milan.

Ferraris, M. (2017), *Postverità e altri enigmi*, Bologna.

Floridi, L. (2020), *Pensare l'infosfera. La filosofia come design concettuale*, Milan.

Floridi, L. (2022), *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, Milan.

Folgheraiter, F. (2015), *Saggi di welfare. Qualità delle relazioni e servizi sociali*, 3rd ed., Trento.

Galli, C. (2022), *Ideologia*, Bologna.

Genovesi, P. (2002), *Utilità della storia. I tempi, gli spazi, gli uomini*, Reggio Emilia.

Huizinga, J. (2012), *La crisi della civiltà*, Milan [*In de schaduwen van morgen. Een diagnose van het geestelijk lijden van onzen tijd*, Haarlem 1935].

Key, V. O. Jr. (1966), *The Responsible Electorate: Rationality in Presidential Voting*, *1936-1960*, with the assistance of M. C. Cummings Jr., Cambridge (MA).

Koselleck, R. (2009), *Storia. La formazione del concetto moderno*, ed. by R. Lista, Bologna [*Geschichte, Historie*, in *Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*, ed. by O. Brunner – W. Conze – R. Koselleck, Band 2, Stuttgart 1975, 647–717].

Lorusso, A. M. (2018), *Postverità*, Rome – Bari.

McIntyre, L. (2019), *Post-Verità*, Novara [*Post-Truth*, Cambridge (MA) 2018].

McPherson, M. – Smith-Lovin, L. – Cook, J.M. (2001), *Birds of a Feather: Homophily in Social Networks*, Annual Review of Sociology 27, 415–44.

Minuti, R. (2017), *Introduzione*, in *Il web e gli studi storici. Guida critica all'uso della rete*, ed. by R. Minuti, 2nd ed., Rome, 11–19.

Nicita, A. (2021), *Il mercato delle verità. Come la disinformazione minaccia la democrazia*, Bologna.

Oxford Languages (2016), *Word of the Year 2016: Post-Truth*, https://languages.oup.com/word-of-the-year/2016/.

Pagnotta, F. (2018), *Introduzione. Il web e la nuova responsabilità ecologica e relazionale dell'umano*, in *Ecologia della Rete. Per una sostenibilità delle relazioni online*, ed. by F. Pagnotta, Trento, 23–34.

Papagno, G. (2000), *Un modello per la storia. Materiale, Attività, Funzione*, Reggio Emilia.

Orecchia, A.M. – Preatoni, D.G. (2022), eds., *Bufale*, fake news, rumors *e post-verità*. *Discipline a confronto*, Sesto San Giovanni (MI).
Quattrociocchi, W. – Vicini, A. (2016), *Misinformation. Guida alla società dell'informazione e della credulità*, Milan.
Quattrociocchi, W. – Vicini, A. (2018), *Liberi di crederci. Informazione, internet e post-verità*, Turin.
Quattrociocchi, W. – Vicini, A. (2023), *Polarizzazioni. Informazioni, opinioni e altri demoni nell'infosfera*, Milan.
Reggiani, N. (2019a), *Papirologia: la cultura scrittoria dell'Egitto greco-romano*, Parma.
Reggiani, N. (2019b), *La papirologia digitale. Prospettiva storico-critica e sviluppi metodologici*, Parma.
Reggiani, N. (2024), *The Artificial Papyrologist at Work*, in *Decoding Cultural Heritage. A Critical Dissection and Taxonomy of Human Creativity through Digital Tools*, ed. by F. Moral-Andrés – E. Merino-Gómez – P. Reviriego, Cham, 123–36.
Riva, G. (2010), *I social network*, Bologna.
Riva, G. (2018), *Fake news. Vivere e sopravvivere in un mondo post-verità*, Bologna.
Segenni, S. 2020, ed., *False notizie… «fake news» e storia romana. Falsificazioni antiche, falsificazioni moderne*, Milan – Florence.
Serughetti, G. (2019), *Introduzione*, in L. McIntyre, *Post-Verità*, Novara, VII–XXIV.
Sorice, M. (2011), *La comunicazione politica*, Rome.
Sorice, M. (2022), *Sociologia dei media. Un'introduzione critica*, 2nd ed., Rome.
Sosin, J. D. (2010), *Digital Papyrology*, "The Stoa Consortium", https://blog.stoa.org/archives/1263.
Strother, J. B. – Fazal, Z. – Ulijn, J. M. (2012), *Information Overload: An International Challenge for Professional Engineers and Technical Communicators*, Hoboken (NJ).
Sunstein, C. R. (2017), *#republic. La democrazia nell'epoca dei social media*, Bologna [*#Republic: Divided Democracy in the Age of Social Media*, Princeton (NJ) 2017].
Tipaldo, G. (2019), *La società della pseudoscienza. Orientarsi tra buone e cattive spiegazioni*, Bologna.
Van Minnen, P. (2009), *The Future of Papyrology*, in *The Oxford Handbook of Papyrology*, ed. by R. S. Bagnall, Oxford – New York, 644–60.
Vitali, S. (2004), *Passato digitale. Le fonti dello storico nell'era del computer*, Milan.
Vernant, J.-P. (2005), *Senza frontiere. Memoria, mito e politica*, Milan [*La traversée des frontières. Entre mythe et politique II*, Paris 2004].
Warlde, C. – Derakhshan H. (2017), *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*, Strasbourg, https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

Monica Berti

Digital Catalogs of Ancient Greek Authors and Works through Papyrological Data

1 Introduction

This paper describes a digital catalog of ancient Greek literature based on the linguistic annotation of references to authors and works in ancient sources. This catalog has been developed by extracting bibliographic data as it was expressed by ancient authors in works that are still extant today thanks to their transmission through manuscripts. The goal of the paper is to analyze the current state of digital papyrology to see if it is possible to extract and annotate this kind of bibliographic data in papyrological texts and contribute to the catalog with further information.

The paper is arranged in four sections. After this short introduction, the second section (*Digital Catalogs and Critical Editing*) discusses the relationship between catalogs and critical editions in a digital environment, while the third section (*Digital Catalogs and Papyrological Data*) and the conclusion investigate linguistic annotation of digital papyrological sources to populate a text-based catalog of ancient Greek authors and works.

2 Digital Catalogs and Critical Editing

Since the past, philology has been always producing collections of sources and bibliographic data related to them. If we consider Classical literature, evidence about ancient libraries – even if generally poor and dispersed – attests the existence of efforts to arrange and catalog textual heritage.[1] One of the most significant examples is certainly represented by the library of Alexandria and the extraordinary work of Alexandrian scholars to collect and catalog textual sources in the Museum. The language of the fragments of the *Pinakes* of Callimachus reveals these efforts as part of the initiatives of the Ptolemaic institution.[2]

This paper has been written thanks to the generous support of the Deutsche Forschungsgemeinschaft (project nr. 434173983: *Textbasierte Extraktion, Analyse und Annotation antiker griechischer Referenzen auf Autoren und Werke*). Whenever possible, online resources cited in this paper have been provided with DOIs and permanent identifiers. Otherwise, the last access was in February 2023.

1 Pfeiffer 1968; Dickey 2007.

2 Berti 2016 with other bibliography. On lists of books and catalogs in the ancient world with a focus on papyrological evidence, see Otranto 2000 and Otranto 2009.

https://doi.org/10.1515/9783111070162-006

Modern philology follows this tradition and shelfs of libraries are rich of catalogs, indices, and other bibliographic tools that are fundamental to preserve information about ancient authors and works. Digital libraries follow this path with new critical issues deriving from their digital nature. If the main goal is always represented by the preservation and transmission of our textual heritage, digital technologies are different from printed ones. If they offer new possibilities for generating new forms of data, they also have to deal with challenges deriving from the transmission of data produced with past technologies including those dating before and after the Gutenberg era. If we narrow our discussion to ancient Greek authors, we dispose of digital catalogs, whose characteristics depend not only on the different periods in which they were conceived, but also on their different purposes and accessibility policies.

Back in the Seventies of last century, the *Thesaurus Linguae Graecae* project began to create a database of authors and works whose texts are collected in its digital library. The result is the so called TLG *Canon*, which is a catalog of Greek literature where authors and works are rendered in the lemmatized Latin form, are assigned three- and four-digit numbers, and are accompanied by bibliographic records of their printed publications.[3] For example, in the TLG *Canon* the Athenian historian Xenophon is identified with `0032` and the *Anabasis* with `0032.006`. Further metadata of the TLG entry – the adjectives *historicus* and *Atheniensis* and the chronological expression *5–4 B.C.* – disambiguates Xenophon and differentiates him from other homonymous authors in the TLG *Canon*. The work title *Anabasis* is accompanied by a reference to the printed edition used for the digitized text in the TLG collection.[4]

The *Perseus Catalog*, which covers Greek, Latin, and other literatures, began in 2005 to ingest and integrate bibliographic metadata of authors and editions produced in more general library systems, in order to create a resource suitable to a digital age. This project started with the assumption that in a true digital environment every text should be a *multitext* reflecting its transmission over the centuries, therefore providing access to as many public domain editions as possible. The *Perseus Catalog* is based on the use of the library-based *Functional Requirements for Bibliographic Records* (FRBR) to organize and represent information about multiple versions of the same work. It also supports the *Canonical Text Services* (CTS) protocol of the *CITE Architecture* to cite and retrieve textual philological data.[5] For example, the Athenian historian Xenophon is identified with `urn:cts:greekLit:tlg0032` and the *Anabasis* with `urn:cts:`

3 For a complete description of the TLG project and its *Canon*, which are available with an institutional or individual subscription at http://stephanus.tlg.uci.edu, see Pantelia 2022, XI–XXXVII.

4 The TLG text of the *Anabasis* is based on the Oxford edition of Edgard C. Marchant. See Pantelia 2022, 798–9.

5 The *Perseus Catalog* is openly accessible at https://catalog.perseus.org. For an introduction and a guide to it, see Babeu 2019. On the *CITE Architecture* see Blackwell and Smith 2019. For a full integration of the catalog, see the *Scaife Viewer*, which is the new reading environment of the *Perseus Digital Library*: https://scaife.perseus.org.

greekLit:tlg0032.tlg006.[6] Different editions of the *Anabasis* are citable with a further element in the CTS URN syntax, like for example urn:cts:greekLit:tlg0032.tlg006.opp-grc9, which is the identifier of the Teubner edition by Karl Hude.[7] Data and metadata of authors, works, and editions are openly accessible in structured XML files of the GitHub repositories of the *Perseus Catalog*.[8]

These short introductions to the TLG *Canon* and the *Perseus Catalog* show how collections of texts and catalogs are always strictly interrelated, as it was in past libraries. Digital technologies offer the possibility to move a step forward and create bibliographic resources based on linguistic information derived from texts collected in digital libraries. Also in traditional indices and catalogs, data about authors and works is derived from textual evidence, but the connection with it is indirectly represented through bibliographic conventions, citations of relevant passages, and commentaries. This is also valid for the first generations of digital libraries and catalogs, which inherit characteristics of traditional libraries. The entries about Xenophon in the TLG *Canon* and in the *Perseus Catalog* provided in the previous paragraphs show the current situation, while digital linguistic resources allow to point directly to textual occurrences, as for example the forms Ξενοφῶν and Ἀναβάσει in the following sentence of the *Deipnosophists* of Athenaeus of Naucratis:

> ἣν παρίστησι γινομένην Ξενοφῶν ὁ καλὸς ἐν τῇ Ἀναβάσει ἐν τῷ παρὰ Σεύθῃ τῷ Θρᾳκὶ συμποσίῳ·
>
> The noble Xenophon in his *Anabasis* describes a dance of this sort that took place at the symposium in the house of Seuthes the Thracian.[9]

In this case, the token Ξενοφῶν can be extracted, cited as urn:cts:greekLit:tlg0008.tlg001.perseus-grc2:1.27@ξενοφῶν[1] (which means that this is the first occurrence of the form Ξενοφῶν in paragraph 27 of book 1 of the *Deipnosophists* of Athenaeus in the edition by Kaibel) and annotated as evidence of the Athenian historian Xenophon, who is identified with urn:cts:greekLit:tlg0032.[10] In the same

6 The entry is accessible at https://catalog.perseus.org/catalog/urn:cite:perseus:author.1499, which embeds the *Perseus* CITE URN urn:cite:perseus:author.1499. As far as ancient Greek authors are concerned, the *Perseus Catalog* ingests numbers from the TLG, which is still the reference system in the community: see Babeu 2019, 55.

7 Editions of the *Anabasis* in the *Perseus Catalog* are available at https://catalog.perseus.org/catalog/urn:cts:greekLit:tlg0032.tlg006. Metadata of the edition by Hude are available at https://catalog.perseus.org/catalog/urn:cts:greekLit:tlg0032.tlg006.opp-grc9 with links to the digitized version in SLUB (*Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden*) and other catalog records in *WorldCat*.

8 https://github.com/PerseusDL/catalog_data.

9 Ath., *Deipn.* I 27 (= 15e). Text and translation by Olson 2006–12.

10 The CTS URN can be part of a URL and therefore web resolvable like in the *Digital Athenaeus* project. See https://www.digitalathenaeus.org/tools/KaibelText/cts_urn_retriever.php?URN=urn:cts:greekLit:tlg0008.tlg001.perseus-grc2:1.27@ξενοφῶν[1] to visualize the token in the text of the *Deipnosophists*. On this re-

passage, the token Ἀναβάσει can be extracted, cited as `urn:cts:greekLit:tlg0008.tlg001.perseus-grc2:1.27@ἀναβάσει[1]`, and annotated as evidence of the work *Anabasis* by Xenophon, which is identified with `urn:cts:greekLit:tlg0032.ath002`.[11]

Figure 1 shows this passage annotated in the *Catalog* of authors and works that has been created starting with data extracted from the *Deipnosophists* of Athenaeus and that is publicly accessible as part of the *Digital Athenaeus* and the *Linked Ancient Greek and Latin* (LAGL) projects.[12]

Fig. 1: *Authors Catalog* of the *Digital Athenaeus* project: Xenophon.

In this figure, it is possible to access all the occurrences of the name of Xenophon and of his works in the text of the *Deipnosophists*. The abbreviation `ath` in the CTS URN `urn:cts:greekLit:tlg0032.ath002` refers to the fact that this is the *Anabasis* cited by Athenaeus, whether he was directly reading the text of Xenophon or finding

source, see Berti 2021, 321–2. On the two reference citation systems of the *Deipnosophists* and for a discussion about citations of Classical sources in a digital environment, see Berti et al. 2016; Berti 2021, 312–20.

11 https://www.digitalathenaeus.org/tools/KaibelText/ctsurnretriever.php?URN=urn:cts:greekLit:tlg0008.tlg001.perseus-grc2:1.27@ἀναβάσει[1].

12 See https://www.digitalathenaeus.org/tools/Catalog/ and https://www.lagl.org, which includes now also the catalog of authors and works extracted from the Lexicon of Harpocration: https://www.lagl.org/tools/harpocration/.

the citation in an intermediate source.[13] In the TLG *Canon* the *Anabasis* is identified with `0032.006`, but this is a reference to the modern edition of the *Anabasis* whose text has been digitized in the TLG.[14] In the *Perseus Catalog* `urn:cts:greekLit:tlg0032.tlg006` is the *work* identifier of the *Anabasis* and is independent of any of its particular manifestations like editions or translations. A *version* identifier is added to differentiate modern editions of the work of Xenophon, as for example `urn:cts:greekLit:tlg0032.tlg006.opp-grc9`, which is the identifier of the Teubner edition of Karl Hude.[15] In the current state, the *Catalog* of the *Digital Athenaeus* project includes also *Wikidata* IDs of ancient Greek authors and works.[16]

Extracting bibliographic information from ancient sources means dealing with many philological questions related to critical editing. In this section of the paper I will provide a few examples, which are of course not exhaustive, but are meant to show different cases that have to be considered when annotating ancient sources in order to produce new digital data.

An example of the complex transmission of information about ancient authors is offered by a passage of book 4 of the *Deipnosophists* (IV 80 = 182c), where Athenaeus presents a discussion about pipes (αὐλοί) which is part of a longer section on musical instruments (IV 75–84 = 174a–185a):

> οἶδα δὲ καὶ ἄλλα γένη αὐλῶν τραγικῶν τε καὶ λυσιῳδικῶν καὶ κιθαριστηρίων, ὧν μνημονεύουσιν Ἔφορός τ᾽ ἐν τοῖς Εὑρήμασι καὶ Εὐφράνωρ ὁ Πυθαγορικὸς ἐν τῷ Περὶ Αὐλῶν, ἔτι δὲ καὶ Ἀριστόξενος καὶ αὐτὸς ἐν τῷ Περὶ Αὐλῶν. ὁ δὲ καλάμινος αὐλὸς τιτύρινος καλεῖται παρὰ τοῖς ἐν Ἰταλίᾳ Δωριεῦσιν, ὡς Ἀρτεμίδωρος ἱστορεῖ ὁ Ἀριστοφάνειος ἐν δευτέρῳ Περὶ Δωρίδος.[17]
>
> I am also familiar with other types of pipes used for tragedy, *lysiēdēs*, and *kithara*-playing, which are mentioned by Ephorus in his *Inventions* and by Euphranor the Pythagorean in his *On Pipes*, as well as by Aristoxenus himself in his *On Pipes*. The Dorians in Italy refer to a reed pipe as a *titurinos*, according to Artemidorus the student of Aristophanes in Book II of *On Doric*: the pipe referred to as a *magadis*.[18]

In this passage, underlined expressions highlight references to authors and titles of their works. A debated case is the name Aristoxenus (Ἀριστόξενος), which has been restituted in the last edition by Douglas Olson on the basis of a comparison with *Deipn.* 14.634d.[19] The oldest witness of the *Deipnosophists* (*Marc. Gr.* 447, 56r) preserves the

13 The number `002` means that this is the second of a total of 11 works of Xenophon cited in the *Deipnosophists*: see Berti 2024.

14 See n. 4.

15 Metadata of editions of the *Anabasis* in the *Perseus Catalog* are accessible at https://catalog.perseus.org/catalog/urn:cts:greekLit:tlg0032.tlg006. On *work* and *version* identifiers in CTS URNs, see Babeu 2019, 55.

16 Berti 2024.

17 Text by Olson 2021.

18 Translation by Olson 2006–12.

19 Olson 2021, 199.

form Ἄλεξις ὤν, which has been kept by Georg Kaibel in the text of his Teubner edition with a note in the critical apparatus about the possible identification with Aristoxenus.[20] Johann Schweighäuser proposed to correct the form of the manuscript with the names Ἀλεξίων or Ἀλέξων, which could identify the grammarian Alexion or the author Alexon of Mindus, who wrote a work on myths according to Diogenes Laertius (I 29).[21]

In the *Catalog* of the *Digital Athenaeus* project, which is based on the text of the edition by Kaibel, the form Ἄλεξις ὤν is citable with the CTS URN `urn:cts:greekLit:tlg0008.tlg001.perseus-grc2:4.80@ἄλεξις[1]-ὤν[2]`, which retrieves the string of text with the two tokens at 4.80 in the Kaibel edition.[22] This string is also annotated both as `urn:cts:greekLit:tlg0088` (Aristoxenus) and as `urn:cts:greekLit:tlg0699` (Alexion) in order to preserve both interpretations of the text.[23] The same double annotation has been used for the string Περὶ Αὐλῶν (`urn:cts:greekLit:tlg0008.tlg001.perseus-grc2:4.80@περὶ[2]-αὐλῶν[3]`), which is the title of the work whose authorship is debated. In this case the CTS URNs used to annotate it are `urn:cts:greekLit:tlg0088.ath010` and `urn:cts:greekLit:tlg0699.ath001`. The linguistic annotation of the title allows to preserve its contextual form in ancient Greek, which is usually hidden in modern editions and catalogs behind generic expressions like *fragmenta* or translations into Latin or other languages.[24]

In the same passage, it is also interesting to see how adjectives are used to further disambiguate and characterize authors, like the forms Πυθαγορικός and Ἀριστοφάνειος for Euphranor and Artemidorus. Both adjectives have been included in the annotation to preserve linguistic information of the text of Athenaeus, as it is possible to see in Figure 2. This language is usually hidden in catalogs, as it is possible to see in the case of Artemidorus in the TLG.[25] The author Euphranor is not yet in the TLG *Canon* and in

20 Kaibel 1887–90, I, 398; III, 573.

21 Schweighäuser 1801–09, II, 667; Canfora 2001, I, 443; III, 1891; IV, 198. On the grammarian Alexion, see Pagani 2015.

22 The edition by Kaibel is accessible in an XML format through the *Perseus Digital Library*. See Berti 2021, 309–11 with links and information on other versions and editions of the text in the repositories of the *Open Greek and Latin* project including the edition of August Meineke. For the new Teubner edition of the *Deipnosophists* by Douglas Olson, see Berti 2022.

23 The two authors are part of the TLG *Canon* and their numbers are ingested in CTS URNs of the *Perseus Catalog*: see Babeu 2019 and Pantelia 2022. Of the two tokens Ἄλεξις ὤν, the first (Ἄλεξις) is also annotated as a personal name (PER), given that the catalog is based on Named Entity annotations, and is accessible through the *Named Entities Digger* and the *Named Entities Concordance* of the *Digital Athenaeus* project: see Berti 2019 and Berti 2024.

24 Cf. Pantelia 2022, 33 and 116.

25 Pantelia 2022, 119. The *Perseus Catalog* has this form in Latin: https://catalog.perseus.org/catalog/urn:cite:perseus:author.209. The inclusion of adjectives and other onomastic elements sometimes generate the problem of not contiguous annotations, like the form Ἀρτεμίδωρος ἱστορεῖ ὁ Ἀριστοφάνειος in *Deipn.* IV 80. On this still not completely solved issue, see Berti forthcoming.

the *Perseus Catalog* and this is the reason why he has been assigned a new identifier: `urn:cts:greekLit:ath0126`. Figure 2 shows the entry of the author in the *Catalog* including his work on pipes (Περὶ Αὐλῶν). Being a new author without an external URN, the abbreviation `ath` has been used to identify him.[26]

Fig. 2: *Authors Catalog* of the *Digital Athenaeus* project: Euphranor

Another example is the use of the same work title for different authors like in *Deipn.* I 22 (= 13c), where Athenaeus lists authors of works on fishing (Ἁλιευτικά):

> οὕτω καὶ ταύτην τὴν τέχνην ἀκριβοῖ μᾶλλον τῶν τοιαῦτα προηγουμένως ἐκδεδωκότων ποιήματα ἢ συγγράμματα, Καίκαλον λέγω τὸν Ἀργεῖον καὶ Νουμήνιον τὸν Ἡρακλεώτην, Παγκράτην τὸν Ἀρκάδα, Ποσειδώνιον τὸν Κορίνθιον καὶ τὸν ὀλίγῳ πρὸ ἡμῶν γενόμενον Ὀππιανὸν τὸν Κίλικα· τοσούτοις γὰρ ἐνετύχομεν ἐποποιοῖς Ἁλιευτικὰ γεγραφόσι· καταλογάδην δὲ τοῖς Σελεύκου τοῦ Ταρσέως καὶ Λεωνίδου τοῦ Βυζαντίου <καὶ Ἀγαθοκλέους τοῦ Ἀτρακίου>.
>
> He is thus more accurate about this art too than are the authors who have published poems or treatises directly concerned with such matters; I am referring to Caecalus of Argos; Numenius of Heracleia; Pancrates of Arcadia; Posidonius of Corinth; and Oppian of Cilicia, who lived shortly before our time. These are all the epic poets we have encountered who have written on fishing, although I have also encountered prose works by Seleucus of Tarsus, Leonidas of Byzantium, and Agathocles of Atrax.[27]

In this case the form Ἁλιευτικά is annotated as a title of eight different authors: `urn:cts:greekLit:tlg2612.ath001` (Caecalus of Argos), `urn:cts:greekLit:tlg0703.ath001` (Numenius of Heraclea), `urn:cts:greekLit:tlg1556.ath002` (Pancrates of Arcadia), `urn:cts:greekLit:ath0242.ath001` (Seleucus of Tarsos), `urn:cts:greekLit:tlg2639.ath001` (Posidonius of Corinth), `urn:cts:greekLit:`

26 For the use of the abbreviation `ath` for work titles, see n. 13.

27 Text and translation by Olson 2006–12.

`tlg0023.ath001` (Oppianus), `urn:cts:greekLit:ath0301.ath001` (Leonidas of Byzantium) and `urn:cts:greekLit:ath0010.ath001` (Agathocles of Atrax).[28] The last author (Agathocles of Atrax) is missing in the Marcianus manuscript of Athenaeus, but has been inserted by Kaibel on a comparison with the text of the *Suda*. The CTS URN, which identifies the string of text of the *Deipnosophists* in the edition of Kaibel (`urn:cts:greekLit:tlg0008.tlg001.perseus-grc2:1.22@ἀγαθοκλέους[1]-ἀτρακίου[1]`) and which is aligned with the identifier of the quoted author (`urn:cts:greekLit:ath0010`), keeps track of the choice of the editor Kaibel to insert this name in the text.

Other interesting possibilities are offered by the alignment between the text and the annotations preserved on the margins of the oldest witness (*Marcianus Graecus* 447) of the *Deipnosophists*.[29] An interesting example is a passage of book 10, where Athenaeus presents a discussion about people who love spending all their time drunk (X 59 = 442b–c):

> Καὶ ὅλα δὲ ἔθνη περὶ μέθας διατρίβοντα μνήμης ἠξίωται· Βαίτων γοῦν ὁ Ἀλεξάνδρου βηματιστὴς ἐν τῷ ἐπιγραφομένῳ Σταθμοὶ τῆς Ἀλεξάνδρου Πορείας καὶ Ἀμύντας ἐν τοῖς Σταθμοῖς τὸ τῶν Ταπύρων ἔθνος φασὶν οὕτω φίλοινον εἶναι ὡς καὶ ἀλείμματι ἄλλῳ μηδενὶ χρῆσθαι ἢ τῷ οἴνῳ. τὰ δ᾽ αὐτὰ ἱστορεῖ καὶ Κτησίας ἐν τῷ Περὶ τῶν κατὰ τὴν Ἀσίαν Φόρων· οὗτος δὲ καὶ δικαιοτάτους αὐτοὺς λέγει εἶναι. Ἁρμόδιος δὲ ὁ Λεπρεάτης ἐν τῷ Περὶ τῶν παρὰ Φιγαλεῦσι Νομίμων φιλοπότας φησὶ γενέσθαι Φιγαλεῖς Μεσσηνίοις ἀστυγείτονας ὄντας καὶ ἀποδημεῖν ἐθισθέντας. Φύλαρχος δ᾽ ἐν ἕκτῃ Βυζαντίους οἰνόφλυγας ὄντας ἐν τοῖς καπηλείοις οἰκεῖν, ἐκμισθώσαντας τοὺς ἑαυτῶν θαλάμους μετὰ τῶν γυναικῶν τοῖς ξένοις, πολεμίας σάλπιγγος οὐδὲ ἐν ὕπνοις ὑπομένοντας ἀκοῦσαι· διὸ καὶ πολεμουμένων ποτὲ αὐτῶν καὶ οὐ προσκαρτερούντων τοῖς τείχεσι Λεωνίδης ὁ στρατηγὸς ἐκέλευσε τὰ καπηλεῖα ἐπὶ τῶν τειχῶν σκηνοπηγεῖν, καὶ μόλις ποτὲ ἐπαύσαντο λιποτακτοῦντες, ὥς φησι Δάμων ἐν τῷ Περὶ Βυζαντίου.[30]

> Whole peoples, moreover, have been thought to deserve being described as spending all their time drunk. Alexander's quartermaster Baiton, for example, in his treatise entitled *Stages of Alexander's Journey*, along with Amyntas in his *Stages*, claim that the Tapyrians like wine so much that they anoint themselves with nothing else. Ctesias in his *On the Tributes Paid throughout Asia* records the same information; he also claims that they are the most honest people in the world. Harmodius of Lepreum in his *On the Customs in Phigaleia* claims that the Phigaleians, whose city is on the Messenian border and who are used to being away from home, like to drink. Phylarchus in Book VI (says) that because the inhabitants of Byzantium guzzle wine, they live in the bars and rent out their own bedrooms, wives and all, to foreigners, and cannot stand to hear a war-trumpet even in their dreams. This is why, when they were being attacked at one point and failed to show any courage in defending their walls, their general Leonides ordered bars to be set up under canopies on top of the walls, and even then they barely stopped deserting their positions, according to Damon in his *On Byzantium*.[31]

28 Other examples are of course represented by the dubious attribution of a work to two or more authors, which often happens in the *Deipnosophists*. Also in this case it is possible to align different annotations to the same string of text.

29 Cipolla 2015.

30 Text by Olson 2020a.

31 Translation by Olson 2006–12.

As it often happens in the text of the *Deipnosophists*, in this passage the work of Phylarchus is not mentioned, but appears only the number of the book from which the citation has been taken. In the margins of the Marcianus manuscript, the anonymus scribe has annotated περὶ τῆς Βυζαντίων οἰνοφλυγίας together with the names of some of the other authors cited in the passage of Athenaeus.[32] This expression is derived from the text of the *Deipnosophists* (Φύλαρχος δ᾽ ἐν ἕκτῃ Βυζαντίους οἰνόφλυγας ὄντας [...]) and is a description of the content of the section of the *Histories* of Phylarchus from which the quotation is derived.[33] Ancient sources, including Athenaeus, are rich of descriptions of the content of ancient works that in modern editions are sometimes treated as forms of ancient titles. If linguistically annotated, the expression of the Marcianus manuscript could be collected under the identifier of the *Histories* of Phylarchus to preserve and retrieve data about the language of bibliographic citations used across the centuries to refer to authors and describe their works that are now lost.[34]

The problem of digital identifiers has not yet been completely solved. As we have seen for other examples, the TLG *Canon* collects *testimonia* and *fragmenta* of Phylarchus under identifiers of their printed editions (`1609.001`, `.002`, `.003`), like the CTS URNs of the *Perseus Catalog*. The *Catalog* of the *Digital Athenaeus* project provides identifiers of fragmentary works and not of modern collections of *fragmenta*. This is the case of the CTS URN `urn:cts:greekLit:tlg1609.ath001`, which identifies the *Histories* of Phylarchus as it is cited in ancient sources and not in modern editions of his fragments. This identifier is also related to one of the authors who preserves his fragments (Athenaeus of Naucratis). The step forward will be to provide identifiers of fragmentary works which will be indenpendent of both modern editions and ancient sources, as it happens for extant sources.[35]

3 Digital Catalogs and Payrological Data

The annotated *Catalog* described in the second section of this paper has been created with data extracted from literary sources preserved through manuscripts. Information about authors and works can be also found in texts coming from other media like papyri and inscriptions.[36] In this section I will analyze the current state of digital papyrological

32 Annotations to *Deipn*. 442b–c in *Marc. Gr*. 447, 198^{v}: Βαίτων | Ἀρμόδιος | Φύλαρχος | περὶ τῆς Βυζαντίων οἰνοφλυγίας. See Cipolla 2015, 121.

33 In the *epitome* of the *Deipnosophists* also the book number of the work of Phylarchus is removed: Φύλαρχος δέ φησι Βυζαντίους οἰνόφλυγας ὄντας [...]. See Olson 2020b, 465.

34 On the characterstics and chronology of *Marc. Gr*. 447 and its annotations, see Cipolla 2015, 1–35.

35 For a detailed discussion of the characteristics of fragments of lost works, which don't exist in themselves but only through the sources that quote them and which therefore require specific identifiers, see Berti 2021, 105–14.

36 Otranto 2000; Otranto 2009; Canali De Rossi 2021.

resources to see where we can find this kind of data and how we can represent it. In order to show that, I will consider the example of the fragmentary historian Hellanicus of Lesbos.

If we look for digital data, the *Canon* of the TLG has an entry about Hellanicus of Lesbos (`0539`) with *testimonia* (`0539.001`) and *fragmenta* (`0539.002` and `0539.003`) extracted from the printed edition of the *Fragmente der griechischen Historiker* (FGrHist) by Felix Jacoby and from a paper by Hans Joachim Mette (1978).[37] As far as the *fragmenta* are concerned, the TLG *Canon* lists four fragments of Hellanicus coming from papyri: PSI X 1173 (= FGrHist 4 F 124b, vol. Ia, p. *6 *Addenda*); P.Oxy. X 1241 (= FGrHist 4 F 189); P.Giss. 307v (= FGrHist 4 F 201 bis, vol. Ia, p. *7 *Addenda*); P.Oxy. XXVI 2442 (= Mette 1978, fr. 133 bis).[38] Other FGrHist fragments of Hellanicus come from papyri, but are not listed in the TLG *Canon*: FGrHist 4 F 19b (= P.Oxy. VIII 1084); FGrHist 4 F 68 (= P.Oxy. XIII 1611); FGrHist 4 F 197bis, vol. Ia, p. *6 *Addenda* (= PSI XIV 1390). Scholarship has also discussed the identification of a fragment of an epic *Atlantias* in P.Oxy. XI 1359, which is probably connected with the *Atlantis* of Hellanicus or with one of his sources.[39]

The texts of these fragments are partly accessible in a digital format. The TLG reproduces the texts of the fragments as they were published in the FGrHist and in the paper of Mette, but the collection is not open and exportable. The projects *Trismegistos* and *Papyri.info* offer access to some of the papyri listed in the previous paragraph.

Trismegistos collects metadata about these papyri: PSI X 1173,[40] P.Oxy. X 1241,[41] P.Giss. 307v,[42] P.Oxy. XXVI 2442,[43] P.Oxy. VIII 1084,[44] P.Oxy. XIII 1611,[45] PSI XIV 1390,[46] P.Oxy. XI 1359.[47]

37 The printed volume of the *Canon* refers also to the edition of the *Fragmenta Historicorum Graecorum* by Karl Müller (`0539.004`): see Pantelia 2022, 373.

38 See Pantelia 2022, XII–XIII on the decision of the TLG project not to include epigraphical and papyrological works, with the exception of literary works preserved on papyri or inscriptions. Cf. Reggiani 2017, 210–1.

39 Robert 1917; Bell 1920, 123. As far as other media are concerned, there is also an inscription among the *testimonia* of Hellanicus (FGrHist 4 T 30 = IG II/III^2 2363), whose digital text is available in *PHI Greek Inscriptions*: https://inscriptions.packhum.org/text/4599: see Berti 2021, 68–75.

40 See https://www.trismegistos.org/text/61611. See also the entry in the online catalog of the *Papiri della Società Italiana* (PSI): http://www.psi-online.it/documents/psi;10;1173.

41 https://www.trismegistos.org/text/63428.

42 https://www.trismegistos.org/text/63250.

43 https://www.trismegistos.org/text/62564.

44 https://www.trismegistos.org/text/59974.

45 https://www.trismegistos.org/text/64211.

46 See https://www.trismegistos.org/tm/detail.php?tm=59773. See also the entry in the online catalog of the *Papiri della Società Italiana* (PSI): http://www.psi-online.it/documents/psi;14;1390.

47 https://www.trismegistos.org/text/60109.

Papyri.info offers the text of the following papyri: P.Oxy. X 1241,[48] P.Oxy. XXVI 2442,[49] P.Oxy. VIII 1084[50] and P.Oxy. XIII 1611.[51] Other four papyri have entries in *Papyri.info*, but without the text: PSI X 1173,[52] P.Giss. 307v,[53] PSI XIV 1390[54] and P.Oxy. XI 1359.[55]

Three papyri whose texts are in *Papyri.info* (P.Oxy. X 1241, P.Oxy. XXVI 2442, and P.Oxy. VIII 1084) and P.Oxy. XIV 1390 are listed in the *Trismegistos* author page about Hellanicus of Lesbos: https://www.trismegistos.org/author/358. This entry provides metadata about Hellanicus and links to external resources such as *Wikipedia*, the manuscript collection of *Pinakes*, the *Perseus Catalog*, the TLG *Canon*, the *Virtual International Authority File* (VIAF), CIRIS and Brill's *Jacoby Online*. As far as the works of Hellanicus are concerned, *Trismegistos* collects four papyri differentiating them between direct attestations and quotations: P.Oxy. VIII 1084 (direct attestation = *Atlantis*), P.Oxy X 1241 (quotation = *opus incertum*), P.Oxy. XXVI 2442 (quotation = *opus incertum*) and PSI XIV 1390 (quotation = *opus incertum*).[56]

P.Oxy. VIII 1084 (https://www.trismegistos.org/text/59974) is dated to the early 2nd century AD and the text has been attributed to the *Atlantis* of Hellanicus of Lesbos. *Trismegistos* offers a detailed description of the papyrus including the attribution to Hellanicus (direct attestation), bibliographic metadata and a link to *Papyri.info* for the text, other metadata, and pictures.[57] This text is also available with a subscription in the reading environment of the *Jacoby Online* of Brill *Scholarly Editions* under FGrHist 4 F 19b and BNJ 4 F 19b.[58] Being a direct attestation, this text doesn't preserve bibliographic data about Hellanicus.

P.Oxy X 1241 (https://www.trismegistos.org/text/63428) is dated to the 2nd century AD and preserves an Alexandrian treatise with catalogs and lists of authorities about mythological and historical information.[59] The name of Hellanicus is restored in column

48 https://papyri.info/dclp/63428.
49 https://papyri.info/dclp/62564.
50 https://papyri.info/dclp/59974.
51 http://papyri.info/dclp/64211.
52 https://papyri.info/dclp/61611.
53 https://papyri.info/dclp/63250.
54 https://papyri.info/dclp/59773.
55 https://papyri.info/dclp/60109.
56 As Mark Depauw has informed me, there are four text types in *Trismegistos*: 1) *Direct attestation*: this means that the text preserves the work of author X; 2) *Quoted*: this means that in the text a work of author X is quoted or referred to; 4) *Commented upon*: this means that a work of author X is the subject of a commentary; 5) *Epitomised*: this means that a work of author X is summarised. In the past there was also 3) *Translated*, but now there is a separate entry in works for each translation. See Berti 2021, 72.
57 https://papyri.info/dclp/59974.
58 See https://scholarlyeditions.brill.com/reader/urn:cts:greekLit:fgrh.0004.bnjo-1-ed-grc:f19b and https://scholarlyeditions.brill.com/reader/urn:cts:greekLit:fgrh.0004.bnjo-2-ed-grc:f19b. The new version of the *Jacoby Online* project has adopted the *CITE Architecture* to cite its contents: see Berti 2021, 63–6.
59 For a discussion on the papyrus in relation to the list of the Alexandrian librarians, see Berti – Costa 2010, 129–34; Murray 2012.

5. The text is available in *Papyri.info*[60] and in the reading environment of Brill's *New Jacoby* under FGrHist 4 F 189 and BNJ 4 F 189.[61] The XML file of the text in *Papyri.info* at l. 3 preserves the restored form `ʽΕλλ<supplied reason="lost">ά</supplied>ν<supplied reason="lost">ι</supplied>κος`.[62] The *Jacoby Online* allows to cite the two fragments of the FGrHist and the BNJ according to the CTS protocol of the *CITE Architecture* (`urn:cts:greekLit:fgrh.0004.bnjo-1-ed-grc:f189` and `urn:cts:greekLit:fgrh.0004.bnjo-2-ed-grc:f189`) and offers a morphological analysis of each token, if available in Morpheus. The *Jacoby Online* allows also to visualize and export the XML file of the Greek text of the fragment published in the FGrHist and in the BNJ with editorial markup in the first case: `ʽΕλ-<note n="544" type="app crit"><p>σιδηρᾶ Wilamowitz</p></note>λάνικος` in the FGrHist and `ʽΕλλάνικος` in the BNJ.[63]

P.Oxy. XXVI 2442 (https://www.trismegistos.org/text/62564) is constituted by several fragments of papyrus dated to the 2nd and 3rd centuries CE with fragments and *scholia* to Pindar that mention the name of Hellanicus (fr. 29, 1–8 = Mette 1978, 7 fr. 133bis = BNJ 4 F 101a).[64] The text is available in *Papyri.info*[65] where the name of Hellanicus is embedded in the string ωνμυποπεριηρουσελλαν\ι/δ[.]..[.][-ca.?-] (fr. 29, 4), which is also marked up in the corresponding XML file:

```
<gap reason="illegible" quantity="5" unit="character"/>
ωνμυποπερ ιηρουσελλαν<add place="above">ι</add>δ
<gap reason="lost" quantity="1" unit="character"/>
<gap reason ="illegible" quantity="2" unit="character"/>
<gap reason= "lost" quantity="1" unit ="character"/>
<gap reason="illegible" quantity="1" unit="character"/>
<gap reason="lost" extent="unknown" unit="character"/>.
```

The critically edited text is available in the *Jacoby Online* under BNJ 4 F 101a in the HTML page and in the corresponding XML file:[66] Ἑλλάνι(κος).

PSI XIV 1390 (www.trismegistos.org/text/59773) is constituted by three fragments dated to the 2nd century AD and contains a *scholion* to Euphorion that mentions the name of Hellanicus (FGrHist 4 F 197bis = BNJ 4 F 197a).[67] The text is available in the *Jaco-*

60 https://papyri.info/dclp/63428.

61 See https://scholarlyeditions.brill.com/reader/urn:cts:greekLit:fgrh.0004.bnjo-1-ed-grc:f189 and https://scholarlyeditions.brill.com/reader/urn:cts:greekLit:fgrh.0004.bnjo-2-ed-grc:f189.

62 On EpiDoc tags for Leiden conventions, see the *Guidelines* at https://epidoc.stoa.org. For their use in *Papyri.info*, see Reggiani 2017, 222–40.

63 XML versions of each fragment of the *Jacoby Online* are separately exportable, but the collection is not open access.

64 Otranto 2000, XXX.

65 https://papyri.info/dclp/62564.

66 https://scholarlyeditions.brill.com/reader/urn:cts:greekLit:fgrh.0004.bnjo-2-ed-grc:f101a.

67 Other metadata about this papyrus is also available at https://relicta.org/cpp/detail.php?CPP=0231, which provides information about the names cited in this papyrus including Hellanicus.

by Online under the *addenda* to FGrHist 4 F 197.[68] In the HTML page the form of the name is edited as Ἑλλάνικο[ς] with a note in the critical apparatus for conjectures, that are preserved in the corresponding XML file:

```
Ἑλλάνικο[ς]<note n="556" type="app_crit">
<p> Ἑλλάνικο[ς] Latte (<hi rend ="italic">Philol.</hi>90, 1935, p. 131)
Ἑλλανίκο[υ] N-V. Aus den Τρωικά?
</p></note>.
```

The text is not yet available in *Papyri.info*.

P.Oxy. XIII 1611 (https://www.trismegistos.org/text/64211) is dated to the early 3rd century AD and contains several fragments of a work on literary criticism with many quotations of lost works. The name of Hellanicus and the title of his work Κτίσεις have been restored in column 2 of fragment 8 (ll. 212–214). The text is available in *Papyri.info*:[69] Ἑλλάνι-]|κος δ᾽ ἐν [ταῖς Ἐθνῶν (?)] | κτίσεσι [-ca.?-] with the corresponding XML encoding:

```
<gap reason="lost" quantity="5" unit= "character"/>
<supplied reason="lost">Ἑλλάνι</supplied>
<lb n="213" break="no"/>κος δ' ἐ<unclear>ν</unclear>
<supplied reason="lost" cert="low">ταῖς Ἐθνῶν </supplied>
<lb n="214"/>κτίσεσ<unclear>ι</unclear>
<gap reason="lost" extent="unknown" unit="character"/>
<milestone rend="paragraphos"unit="undefined"/>.
```

The text is also available in the *Jacoby Online* under FGrHist 4 F 68 and BNJ 4 F 68:[70] Ἑλλάνι]κος δ᾽ ἐν [ταῖς Ἐθνῶν (?)| Κτίσεσι []| in FGrHist and Ἑλλάνι]κος δ᾽ ἐν [ταῖς Ἐθνῶν (?)| Κτίσεσι [**]| in the BNJ. The XML file of the FGrHist fragment doesn't include EpiDoc tags, that are present in the XML file of the BNJ fragment:

```
<hi rend="bold">]| τοι συμ[</hi>]|βιων π̣ [… Ἑλλάνι]κος δ᾽ ἐν [ταῖς
Ἐθνῶν (?)|<note n="234" type="app_crit">
<p>erg von Allen</p></note>Κτίσεσι [<hi rend="bold">]|.
```

68 https://scholarlyeditions.brill.com/reader/urn:cts:greekLit:fgrh.0004.bnjo-1-ed-grc:f197/?right=bnjo-1-comm1-eng.

69 https://papyri.info/dclp/64211.

70 See https://scholarlyeditions.brill.com/reader/urn:cts:greekLit:fgrh.0004.bnjo-1-ed-grc:f68 and https://scholarlyeditions.brill.com/reader/urn:cts:greekLit:fgrh.0004.bnjo-2-ed-grc:f68.

4 Conclusion

The examples analyzed in the third section of this paper, even if limited to one author and to a small group of texts, show the complexities of papyrological sources, which are quite challenging for extracting ancient references to authors and works. In our case, the fragmented nature of the papyri and the editorial work of philologists make difficult to extract the ancient Greek occurrences of the name of Hellanicus and of the descriptions of his works.

Recent scholarship has been experimenting with linguistic annotation of documentary and literary papyri and has been producing first significant results for the morphological analysis and lemmatization.[71] Nevertheless, many other issues have still to be addressed and solved, such as the variety of papyrological data at our disposal, the stratification of editorial interventions on texts whose nature is generally very fragmented, and different levels of accessibility to ancient sources provided by digital projects.[72] This situation is inevitably depending on the complexities of historical languages and on the long history of philology applied to papyrological evidence, which all result in a slow move to satisfactory digital data.[73]

As far as the linguistic component is concerned, *Trismegistos Words* offers the possibility to search for lemmata and their inflected forms extracted from the XML files of Greek papyri available in *Papyri.info*.[74] As we are warned in the *About* page of this project, coverage and accuracy of data are affected by the limits of the trained algorithm, which still generates mistakes in the part-of-speech/morphology tagging, especially for damaged words and sections. Moreover, still missing are numeric identifiers of lemmata, of occurrences of the words, and of the morphological analysis, given that "changes in the underlying text" has to be properly addressed.[75] Solutions to all these aspects are currently the research focus of linguistics applied to papyrological texts, but time and manual interventions are certainly still needed to cover the impressive amount of data that has been transmitted to us from the past.

Extracting references to ancient Greek authors and to their works is an even more demanding task, because it involves further analyses of proper names and of the content of papyrological sources. Named Entity recognition and disambiguation are part of

71 Reggiani 2017, 178–89; Celano 2018; Vierros – Henriksson 2018; Keersmaekers 2020; Vierros 2021; Keersmaekers – Depauw 2022; Vierros – Yordanova 2022.

72 Different accessibility policies of *Trismegistos*, *Papyri.info*, and the *Jacoby Online* described in section 3 are quite exemplary.

73 Cf. Reggiani 2022.

74 See https://www.trismegistos.org/words. Other projects producing linguistic data from Greek papyri are *Morphologically Annotated and Lemmatized Papyri Corpus* (MALP) and *Sematia*: see Celano 2018 and Vierros 2018.

75 https://www.trismegistos.org/words/about.php.

this task in order to produce a first level of annotation and to extract explicit forms containing also names of ancient authors and titles of their works, or at least part of them. Named Entities have been extracted and annotated in the *Catalog* of the *Digital Athenaeus* project described in section 2 of this paper and have been used to disambiguate authors and works.[76] In this regard, *Trismegistos People* covers personal names of non-royal individuals living in Egypt in documentary texts and names from Greek inscriptions of the Ptolemaic period, but more work on the part of the community still needs to be added to this effort to cover also literary papyri.[77]

All these comments are not meant to underestimate the quality of these resources, but to show the current state of the art of papyrological data related to ancient bibliographic references. Publications, workshops, and projects are producing more and more research questions to find proper solutions for future results in this field of studies, where it is necessary to separate different levels of historical, philological, and linguistic analyses of texts, and to conceive suitable citation systems and exchange policies in the spirit of the Linked Open Data (LOD) initiative for the ancient world.[78]

Bibliography

Babeu, A. (2019), *The Perseus Catalog: of FRBR, Finding Aids, Linked Data, and Open Greek and Latin*, in *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, ed. by M. Berti, Berlin – Boston, 53–72. [https://doi.org/10.1515/9783110599572-005]

Bell, H. I. (1920), *Bibliography: Graeco-Roman Egypt A. Papyri (1915–1919)*, JEA 6, 119–46.

Berti, M. (2016), *Greek and Roman Libraries in the Hellenistic Age*, in *The Dead Sea Scrolls at Qumran and the Concept of a Library*, ed. by S. White Crawford – C. Wassen, Leiden, 31–54. [https://doi.org/10.1163/9789004305069_005]

Berti, M. (2019), *Named Entity Annotation for Ancient Greek with INCEpTION*, in *Proceedings of CLARIN Annual Conference 2019*, ed. by K. Simov – M. Eskevich, Leipzig, 1–4.

Berti, M. (2021), *Digital Editions of Historical Fragmentary Texts*, Heidelberg. DOI: 10.11588/propylaeum.898.

Berti, M. (2022), Review of S. Douglas Olson (ed.) *Athenaeus Naucratites. Deipnosophistae*. Vol. III.A–B, LIBRI VIII–XI, Berlin – Boston 2020. S. Douglas Olson (ed.) *Athenaeus Naucratites. Deipnosophistae*. Vol. IV.A–B, LIBRI XII–XV, Berlin – Boston 2019, ExClass 26, https://doi.org/10.33776/ec.v26.7422.

Berti, M. (2024), *Digital Canons and Catalogs of Fragmentary Literature*, in *Fragmente einer fragmentierten Welt. Zur Problematik des Umgangs mit Fragmenten in der gegenwärtigen klassisch-philologischen Forschung*, hrsg. von F. Neuerburg – T. Tsiampokalos – P. Wozniczka, Berlin – Boston (in press).

Berti, M. (forthcoming), *Linguistic Annotation for a Catalog of Ancient Greek Authors and Works*, in *Proceedings of the 10th Intenational Colloquium on Ancient Greek Linguistics* (ICAGL).

76 See n. 23.

77 See https://www.trismegistos.org/ref. For example, occurrences of the Greek name Ἑλλάνικος can be found at www.trismegistos.org/namvar/3015, which doesn't include literary papyri and therefore the author with this name. On Named Entities in the *Trismegistos* project, see Broux – Depauw 2015.

78 Cf. Celano 2018, 139; Cayless 2019.

Berti, M. – Blackwell, C. W. – Daniels, M. – Strickland, S. – Vincent-Dobbins, K. (2016), *Documenting Homeric Text-Reuse in the* Deipnosophistae *of Athenaeus of Naucratis*, in *Digital Approaches and the Ancient World*, ed. by G. Bodard – Y. Broux – S. Tarte (BICS Themed Issue 59.2), 121–39. [https://doi.org/10.1111/j.2041-5370.2016.12042.x]

Berti, M. – Costa, V. (2010), *La Biblioteca di Alessandria. Storia di un paradiso perduto*, Tivoli.

Blackwell, C. W. – Smith, N. (2019), *The CITE Architecture: a Conceptual and Practical Overview*, in *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, ed. by M. Berti, Berlin – Boston, 73–94. [https://doi.org/10.1515/9783110599572-006].

Broux, Y. – Depauw, M. (2015), *Developing Onomastic Gazetteers and Prosopographies for the Ancient World Through Named Entity Recognition and Graph Visualization. Some Examples from Trismegistos People*, in *Social Informatics. SocInfo 2014. Lecture Notes in Computer Science*, vol 8852, ed. by L. Aiello – D. McFarland (eds.), Berlin, 304–13. [https://doi.org/10.1007/978-3-319-15168-7_38]

Canali De Rossi, F. (2021), *Le iscrizioni degli antichi autori greci e latini*, Rome.

Canfora, L. (2001), *Ateneo. I Deipnosofisti. I dotti a banchetto*, Rome.

Cayless, H. A. (2019), *Sustaining Linked Ancient World Data*, in *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, ed. by M. Berti, Berlin – Boston, 35–50. [https://doi.org/10.1515/9783110599572-004]

Celano, G. G. A. (2018), *An Automatic Morphological Annotation and Lemmatization for the IDP Papyri*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 139–148. [https://doi.org/10.1515/9783110547450-008]

Cipolla, P. (2015), *Marginalia in Athenaeum. Lemmi, scoli e note di lettura del codice Marc. Gr. 447 dei* Deipnosofisti. Amsterdam.

Dickey, E. (2007), *Ancient Greek Scholarship. A Guide to Finding, Reading, and Understanding Scholia, Commentaries, Lexica, and Grammatical Treatises, from Their Beginnings to the Byzantine Period*, Oxford.

Kaibel G. (1887–90), *Athenaei Naucratitae Dipnosophistarum libri 15*, Lipsiae.

Keersmaekers, A. (2020), *Creating a Richly Annotated Corpus of Papyrological Greek. The Possibilities of Natural Language Processing Approaches to a Highly Inflected Historical Language*, Digital Scholarship in the Humanities 35, 67–82. [https://doi.org/10.1093/llc/fqz004]

Keersmaekers, A. – Depauw, M. (2022), *Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri*, in *Digital Text Analysis of Greek and Latin Sources; Methods, Tools, Perspectives* (Classics@ 20), ed. by S. Chronopoulos – F. K. Maier – A. Novokhatko, Washington (DC).

Mette, H. J. (1978), *Die «Kleinen» griechischen Historiker heute*, Lustrum 21, 5–43.

Murray, J. (2012), *Burned after Reading. The So-Called List of Alexandrian Librarians in P. Oxy. X 1241*, Aitia 2, https://doi.org/10.4000/aitia.544.

Olson, S. D. (2006–12), *Athenaeus. The Learned Banqueters*, Cambridge (MA).

Olson, S. D. (2020a), *Athenaeus Naucratites. Deipnosophistae*, III.A (*Libri VIII–XI*), Berlin – Boston.

Olson, S. D. (2020b), *Athenaeus Naucratites. Deipnosophistae*, III.A. (*Libri VIII–XI: Epitome*), Berlin – Boston.

Olson, S. D. (2021), *Athenaeus Naucratites. Deipnosophistae*, II.A (*Libri III.74–VII*), Berlin – Boston.

Otranto, R. (2000), *Antiche liste di libri su papiro*, Rome.

Otranto, R. (2009), *Liste di libri su papiro. Tra conservazione e perdita*, Atene e Roma 1–2, 13–32.

Pagani, L. (2015), *Alexion Cholus*, in *Lexicon of Greek Grammarians of Antiquity*, ed. by F. Montanari – F. Montana – L. Pagani, Leiden – Boston, https://doi.org/10.1163/2451-9278.

Pantelia, M. C. (2022), *Thesaurus Linguae Graecae. A Bibliographic Guide to the Canon of Greek Authors and Works*, Oakland (CA).

Pfeiffer, R. (1968), *History of Classical Scholarship. From the Beginning to the End of the Hellenistic Age*, Oxford.

Reggiani, N. (2017), *Digital Papyrology I. Methods, Tools and Trends*, Berlin – Boston. [https://doi.org/10.1515/9783110547474]

Reggiani, N. (2022), *The Digital Edition of Ancient Sources as a Further Step in the Textual Transmission*, in *Digital Text Analysis of Greek and Latin Sources; Methods, Tools, Perspectives* (Classics@ 20), ed. by S. Chronopoulos – F. K. Maier – A. Novokhatko, Washington (DC).

Robert, C. (1917), *Eine epische* Atlantias, Hermes 52, 477–9.

Schweighäuser, J. (1801–09), *Animadversiones in Athenaei Deipnosophistas*, Argentorati.

Vierros, M. (2018), *Linguistic Annotation of the Digital Papyrological Corpus: Sematia*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 105–18. [https://doi.org/10.1515/9783110547450-006]

Vierros, M. – Henriksson, E. (2021), *PapyGreek Treebanks: A Dataset of Linguistically Annotated Greek Documentary Papyri*, Journal of Open Humanities Data 7.26, https://doi.org/10.5334/johd.55.

Vierros, M. – Yordanova, P. V. (2022), *Querying Syntactic Constructions in Ancient Greek Parsed Corpora. A Case Study on the Genitive Absolute in Literature and Documentary Papyri*, in *Digital Text Analysis of Greek and Latin Sources; Methods, Tools, Perspectives* (Classics@ 20), ed. by S. Chronopoulos – F. K. Maier – A. Novokhatko, Washington (DC).

Mark Depauw

Why Not to Choose XML, Or the Importance of Identifiers

The title of this paper is provocative. And it is meant to be so. Over the last twenty years or so, I have been involved in many Digital Humanities projects dealing with the Ancient World, and XML is omnipresent. In some environments, particularly those dealing with inscriptions and papyri, it feels as if XML is a panacea for every ailment. Almost a dogma, it is better to ask absolution for your sin before presenting a project and admitting that you are not using XML. Yet I think that XML is not the only way to structure data and not the best answer to all problems faced when developing research infrastructures. Relational databases – unsurprisingly perhaps the system I use – are a real alternative that should be considered. And in both systems good identifiers are crucial.

I am not the first one to point out that there are also downsides to XML. A Google search finds many other discussions, pleas or even tirades. The main argument often is that XML is heavy, hardly legible to the human eye, "with its sharp, pointy angle brackets, jabbing you directly in your ever-lovin' eyeballs".[1] Apart from aesthetic concerns (and perhaps related to them) the system of tags enclosing relevant information is undeniably cumbersome. On top of that, XML's strict hierarchical structure can lead to unnecessary reduplication of tags. It is therefore not surprising that successful XML-projects such as the Papyrological Navigator [PN] have limited their annotation to aspects of text reconstruction and layout (e.g., lacunae, abbreviations, erroneous omissions, superfluous elements; texts structure in columns and lines).[2] On the basis of the old beta code of the DDbDP, the PN produced a versatile and lightweight XML edition which has proved an indispensable starting point for many satellite projects. These then build new infrastructures on top of a time-stamped copy of the PN edition.

About a decade ago, I was involved in such as project. Alek Keersmaekers started from the XML-text of Greek papyri in the PN (state of September 2016).[3] He applied machine learning to tokenize, lemmatize and part-of-speech-tag the corpus. His final goal was to also parse the texts automatically and then study shifts in Greek grammar based on the corpus. But convicted that his preliminary results were already very useful for the papyrological community, Alek and I developed Trismegistos Words, a website presenting the morphological analysis and lemmatisation.[4] To produce statistics and elaborate but intuitive search possibilities, Trismegistos turned to what it knew best: a MySQL relational database structure. All information in the XML was converted

1 https://blog.codinghorror.com/xml-the-angle-bracket-tax. All hyperlinks last accessed on 7.3.2024.
2 https://papyri.info.
3 See Keersmaekers – Depauw 2020.
4 https://www.trismegistos.org/words.

https://doi.org/10.1515/9783111070162-007

accordingly, and eventually a search interface was launched. It allowed to quantify the chronological spread of certain words, offering filters based on cases, numbers, or all other morphological elements. For TM Texts – and with kind permission of the PN –, we also reconstructed the text on the basis of the MySQL database and presented it line-by-line or continuously according to the user's preference. We also threw in the morphological analysis, which is particularly useful for people whose knowledge of Greek is limited – a growing number amongst students and laypeople. But we also went a step further and used this database of almost 4.5 million words as an anchor point for much of the Trismegistos information that we had distilled from the full text in earlier projects. Where-ever possible, textual references to people, places, dates, formulae and others were connected to the full text by using the individual TM WordRef ID, which we had added when transforming the text into a database. This improved upon the already existing connection to the full text through the identification of fragment, section, column, and line 'numbers'. For although these elements could in theory also be structured in a perfectly logical way, experience teaches that the system is often tweaked by editors to deal with idiosyncratic aberrations. An additional complication is that some papyri are so broad that the same element occurs multiple times in a single line.

This brings me to a crucial point: IDs. Trismegistos has long been active in this domain, and assigning persistent identifiers is fundamental to what we do.[5] The Trismegistos Text ID (mostly just called TM ID) is used by projects to identify which texts they have in their collection, and to connect to websites with related information. The Papyrological Navigator uses it to tie together what historically are its constituent elements: the metadata (from the Heidelberger Gesamtverzeichnis – HGV), the full text (from the Duke Databank of Documentary Papyri – DDbDP), and the images (from the Advanced Papyrological Information System – APIS). But the TM ID has a rather low granularity: a text consists of many elements such as sections, columns, lines and of course individual words, and no IDs for these are used or provided by the PN. These ID proxies (sections, columns, ...) are also not persistent. Different editions may use different systems, lines may be added or turn out to be no ink but an artefact of the photograph, etc. But above all, the proxies are considered as meaningful by humans. Line 2 follows line 1 and column A precedes column B: if this is not the case the feeling is that something is wrong. As a result, editors feel obliged to change the line numbers after a rearrangement of the fragments, making line 1 of papyrus # no longer line 1 of papyrus # in a previous edition.

What we need therefore is a stable way to refer to elements inside a text, identifiers with a lower granularity. These should not even be connected to the TM number, as this would again lead to complications if a fragment were reassigned to a different text (with a different TM ID). It is essential that this number (or any other string) is meaningless. One could think of a UU_ID for example, although this is impractical for human use and

5 Depauw 2018.

therefore impedes manual data exchange and connections between projects. A simple number is an equally good possibility, although it should be very clearly communicated that it is not a 'serial' number and that any perceived cardinality is imaginary: ID 105 does not necessarily follow ID 104 and will remain ID 105 even if it is transferred to an environment of numbers in the 100.000s, for example. The TM WordRef IDs should be seen as an attempt to introduce such as persistent identifier for each word in papyrological texts.

In an XML environment such identifiers are normally absent, and information on the location of textual elements is as a rule implicit. The order of words, for example, is reflected in their position in the xml-file. Structural layout elements such as columns and lines are in many XML-datasets pointed out by specific tags, but – as stated above – these do not act as identifiers, but as meaningful elements. This often has the perverse effect of causing projects which are not necessarily interested in the full text, but only in specific elements in it, to include the text of the entire document. This in turn leads to the multiplicity of text editions, causing not only reduplication of effort, but also confusion. Which text is the best? Should text that is not the core business of some projects also be updated, or is this only necessary for the mother dataset, e.g., the Papyrological Navigator?

In my mind there are two important steps to be taken towards a solution. Firstly, stable identifiers should be introduced for individual words in the environment where the canonical version of the text lies, for papyri preferably – of course – the Papyrological Navigator. This may seem a simple and uncomplicated addition, but it is in fact a very complex innovation. To begin with, assigning persistent identifiers implies being able to keep track of changes. References to a word that has been re-read should be redirected to the new reading, even if this consists of two or more words. While this is certainly possible in XML, the texts risk to be clogged with annotations even further. In a database, this kind of administration is somewhat easier.

The second step towards greater interoperability would be for projects to stop re-editing texts with the sole purpose of annotating specific aspects of it. It is much leaner to focus on those words that the project is concerned with. These are the ones that should be connected to the full edition of the text in the PN, through the identifier. Trismegistos is currently exploring how this could be done in a new project called NIKAW.[6] The aim is to reconstruct networks of entities extracted from the full text of (literary) classical authors present in corpora. As a first step, we have looked at LASLA/LILA,[7] where a fully tokenized text with individual word IDs is available in tabular form. Through Named Entity Extraction [NER], matching with the TM gazetteers and manual checks, we have and are still compiling attestations of people (and places, while we are

6 For NIKAW [Networks of Ideas and Knowledge in the Ancient World], see https://research.kuleuven.be/portal/en/project/3H230094.

7 Fantoli – Passarotti – Mambrini *et al.* 2022.

at it). Each of these is connected to the full text through the LILA-token ID or IDs. We do not need to worry about the entire text, but we can pull it in whenever we think it is necessary. For NIKAW, we will proceed to tackle other full text corpora such as the new Greek corpus GLAUx in the same way.[8] There is only one prerequisite: each word should have a stable ID.

This will not be the final step. An edition is of course only an interpretation, and it would be more objective to connect to the papyrus itself, or at least a visual representation (and interpretation?) of it. IIIF certainly has a role to play here, but the community should be careful to centralize this effort and avoid annotating different visual representations of the same object. Otherwise, a similar problem as for XML may occur.

Bibliography

Depauw, M. (2018), *Trismegistos. Optimizing Interoperability for Texts from the Ancient World*, in *Crossing Experiences in Digital Epigraphy. From Practice to Discipline*, ed. by A. De Santis – I. Rossi, Berlin – Boston, 191–9.

Fantoli, M. –Passarotti, M. – Mambrini, F. – Moretti, G. – Ruffolo, P. (2022), *Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin*, in *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, ed. by T. Declerck *et al.*, Marseille, 26–34.

Keersmaekers, A. (2021), *The GLAUx Corpus: Methodological Issues in Designing a Long-Term, Diverse, Multi-Layered Corpus of Ancient Greek*, in *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change*, ed. by N. Tahmasebi – A. Jatowt – Y. Xu – S. Hengchen – S. Montariol – H. Dubossarsky, Stroudsburg (PA), 39–50. [https://aclanthology.org/2021.lchange-1.6]

Keersmaekers, A. – Depauw, M. (2020), *Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri*, in *Digital Methods of Analysing and Reconstructing Ancient Greek and Latin Texts*, ed. by A. Novokhatko – F. K. Maier (Classics@ 20), https://classics-at.chs.harvard.edu/bringing-together-linguistics-and-social-history-in-automated-text-analysis-of-greek-papyri.

8 Keersmaekers 2021.

Part II: **Text Encoding, Editing, and Linguistic Applications**

Marie-Pierre Chaufray — Lorenzo Uggetti

GESHAEM and the Challenge of Encoding Greek and Demotic Papyri

1 GESHAEM: the Graeco-Egyptian State – Hellenistic Archives from Egyptian Mummies

The project GESHAEM (The Graeco-Egyptian State – Hellenistic Archives from Egyptian Mummies), directed by Marie-Pierre Chaufray at the Ausonius Institute in Bordeaux, aims to improve our knowledge of the administration of the Fayyum during the first century of Ptolemaic rule by studying texts that have long been neglected: the papyri of the Jouguet Collection in Paris. This paper will not present the historical issues concerning the documents, but, after a short description of the context and the object of study, the questions raised by the encoding and some of the choices made for the website database, implemented by Nathalie Prévôt, the software engineer for digital humanities at the Ausonius Institute in Bordeaux[1].

The Jouguet Collection is a bilingual corpus, with texts in Greek and in Egyptian demotic. Most of the texts date from the 3rd century BC (with a few texts from the 2nd century). Administrative documents prevail, such as contracts, letters, petitions etc., which belonged to the archives of officials at various levels of the administration. A few literary texts have also been found[2]: fragments from Menander, Homer and Euripides and, in demotic, one fragment of a wisdom text. Greek documents have been published in the first volumes of P.Lille[3], and then P.Sorb.;[4] the *enteuxeis* (Greek petitions) were gathered in a special volume[5]. Some accounts have been published in the P.Count[6] and other texts have been published separately. In total, around 330 Greek documents have been published and 11 literary texts. The demotic texts have been published in 3

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 758907). Many thanks to Faye Wills, MA BA (Hons) PGCE MCCT, PhD candidate at LASAR Canterbury Christ Church University, for checking and correcting our English. Any shortcoming remains our responsibility. The first sub-chapter, *GESHAEM: the Graeco-Egyptian State – Hellenistic Archives from Egyptian Mummies*, has been authored by M.-P. Chaufray, the second one, *The Challenge of Encoding Greek and Demotic Papyri*, by L. Uggetti.

1 https://geshaem.huma-num.fr.

2 Uggetti 2022a, 326–7.

3 Jouguet 1907–28 (P.Lille I); Lesquier 1912 (P.Lille II).

4 Cadell 1966 (P.Sorb. I); Cadell – Clarysse – Robic 2011 (P.Sorb. III).

5 Guéraud 1931 (P.Enteux.).

6 Clarysse – Thompson 2006.

https://doi.org/10.1515/9783111070162-008

volumes of P.LilleDem.[7] and in one volume of P.Sorb.[8]. Other documents have been published separately in periodicals or collective works. In total, around 500 papyri of the Jouguet Collection have been published.

Given that there are today around 1000 inventory numbers of the Jouguet Collection in the book register of the Sorbonne, and that one inventory number often contains several fragments, a lot remains to be done on the collection. The main reason for this rather low level of publication is the state of the texts, which is linked to their provenance. The Jouguet Collection comes mostly from Pierre Jouguet's excavations in the Fayyum which took place in the winters of 1901 and 1902 at the southwestern sites of Medinet Ghôran and Medinet Nehas, the ancient Magdôla[9]. In the necropoleis, Jouguet discovered hundreds of mummy decorations called cartonnages. Most of them date from the 3rd and 2nd centuries BC. The majority of these pieces of cartonnages were made of papyri, and they were destroyed soon after their discovery, in order to take out the papyri. Only 20 or so pieces of cartonnages remain today in good overall condition. Part of the project GESHAEM is concerned with the restoration and study of these objects, but the major part of the project is the study of the texts, among which some new texts which have been extracted from the pieces of cartonnages that existed within the collection in very poor condition. Some had been partially destroyed and the plaster had been almost completely removed from their surface: it was even difficult to recognise the different elements of the original cartonnage[10].

The result of the extraction gave 438 new inventory numbers, which were photographed by Adam Bülow-Jacobsen both in colour and in infrared. Like in the rest of the Jouguet Collection, the texts are very fragmentary. One major goal of GESHAEM was therefore to develop a computerised tool to help in the reconstruction of the texts by automatically proposing suggestions of pairing papyri. This part of the project was completed by Antoine Pirrone in 2022[11], at the Laboratoire Bordelais de Recherche en Informatique (LaBRI)[12]. Given the state of the fragments, a discussion on associating some metadata to the images to help the computer with the matching of fragments took place at the beginning of the project. This information has been entered into the project database in XML, and some encoding issues came up.

The other encoding issues concern the content of the texts that will be published within GESHAEM. The corpus of bilingual surety contracts is planned to be displayed online. Part of this corpus was published by Françoise de Cenival in 1973[13], but she pub-

7 Sottas 1921 (P.LilleDem. I); De Cenival 1973 (P.LilleDem. II); De Cenival 1984 (P.LilleDem. III).

8 Chaufray – Wackenier 2016 (P.Sorb. IV). The P.Sorb. III 76, 78, 81, 83 and 85 also bear demotic texts with Greek subscriptions, except for the last one: Cadell – Clarysse – Robic 2011, 57–71.

9 Jouguet 1901; Jouguet 1902; Jouguet – Lefebvre 1902.

10 Uggetti 2022b, 989–93 and 997–1000.

11 Pirrone 2022.

12 https://www.labri.fr.

13 See above, n. 7.

lished only the demotic recto of the documents and not the Greek docket on the verso, which gives a short summary of the demotic text. Besides, many more fragments have been discovered in the Jouguet Collection since de Cenival's publication, which makes a corpus of around 200 documents, which Marie-Pierre Chaufray and Willy Clarysse are currently working on. Specific aspects of these documents are the legal clauses of the demotic contract, which will be highlighted in the online publication. Another corpus of demotic tax-accounts coming from Magdôla will also be published and displayed online. These documents are long lists of taxes paid by individuals, often with names in one column and amounts of grains or money in another column. The online publication would like to help the reader not only find information on the people and the villages mentioned in those texts, but also on making calculations and statistics on these documents.

2 The Challenge of Encoding Greek and Demotic Papyri

Each document is represented digitally by one XML file, which is univocally connected to four photographs: two in colour and two in infrared, so as to have a comprehensive photo coverage of both recto and verso. Every XML file name is formed by a specific identification number preceded by the letter g, standing for GESHAEM, which does not correspond to Sorbonne inventory numbers[14]: the project focuses on a part of the Sorbonne collection and not on the whole of it, so adopting inventory numbers as file names would have brought about gaps in the numbering.

The definition of recto and verso of a papyrus is nested in the file section concerning the textual content, that is under the root `text`, within the child `body`: a first element `division` introduces the encoding of the whole text, whereas a second subchild element `division` represents the single papyrus thanks to the attribute value `fragment`. Both sides are defined by a further element `division` bearing the initials of either "recto" or "verso" as an attribute value: an `xml:id` is assigned to each of them, formed by the XML file name followed by the ending `-recto` or `-verso`[15].

For each side, a correspondence with this `xml:id` is established in the file section `facsimile`, which lists the pictures of the described object, thanks to an attribute

14 For instance, Inv. Sorb. 779 and 1257, joined together, bear the number `g335`.

15 As an example:

```
<text>
<body>
<div type="edition">
<div n="1" type="textpart" subtype="fragment">
<div n="r" type="textpart" ana="#perfibral" xml:id="g335-recto">
```

`correspondence` associated to the child element `surface`. The subchild element `graphic` links each single photo to the Uniform Resource Locator (URL) of its digital repository. A further subchild element `description` provides a name for the images formed by their Sorbonne inventory numbers, followed by acronyms which help recognise whether they are colour or infrared images, either of recto or verso[16]. This way, the XML file itself links unambiguously the parts of text on each side of a papyrus to the metadata describing the pictures. The repository chosen is NAKALA, a service created by the Research Infrastructure Huma-Num specifically dedicated to humanities and social sciences[17]: it ensures long-term data preservation, is in accordance to the International Image Interoperability Framework[18], and respects eco-design criteria, as guaranteed by the GreenWeb certified index[19].

In order to add metadata which might be helpful for the automatic matching of fragments, specific attention has been given to the encoding of the physical description of the papyri[20]: in particular, the *kollêsis*, that is the narrow area where two sheets are pasted together, in order to make a roll[21]. When a text is drafted along the fibres on the internal side of a roll, such overlaps run perpendicularly to the written lines: if two or more fragments show the same approximate position for a *kollêsis*, they might belong to the same vertical section of a papyrus, even if there is a gap between them which does not allow a direct join. This is particularly useful when reconstructing documents with a high and narrow format, such as the approximately 200 surety contracts within the Jouguet Collection[22]. It is often easy to detect *kollêseis* when looking at the original papyri, but sometimes it is difficult to see them on photographs (for example, when they are very close to the borders): therefore, it is better to point them out clearly in the XML files, nested in the physical description of the support, next to the overall dimensions of each papyrus. They are encoded as `referencing string`, with two attributes: `type`, using the value `kollesis` to introduce the concept itself within the file, and `number`, in case of multiple ones on the same fragment. Their distance has been measured in

16 For instance:

```
<facsimile>
<surface corresp="#g335-recto">
<graphic url="https://www.nakala.fr/iiif/11280/084adedb">
<desc type="R_IR">0779_1257_r_IR</desc>
<desc type="view">Image infra-rouge du recto</desc>
<desc type="copyright">Adam Bülow-Jacobsen</desc>
```

17 https://www.nakala.fr; https://documentation.huma-num.fr/nakala-guide-de-description.

18 https://iiif.io.

19 With a score of 90 over 100: https://green-web.fr/index/www-nakala-fr/?lang=en.

20 Analysing some sample files, other papyrological databases, like the Duke Data Bank of Documentary Papyri (https://papyri.info) and the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (https://aquila.zaw.uni-heidelberg.de), have apparently made different choices.

21 Bülow-Jacobsen 2009, 19–21.

22 New edition forthcoming by Marie-Pierre Chaufray and Willy Clarysse: see *supra*, §1.

centimetres from the right border of the papyrus, so that fragments showing similar distances might be grouped together[23].

In order to avoid false joins between fragments, the presence of margins has been highlighted. Their preservation is often accidental and might also be partial: as a consequence, the choice has been made to indicate their position, but not to estimate their width. First, an element `layout` has been used to create a correspondence with the `xml:id` identifying the different sides of the object. Then, depending on the number of margins actually preserved (from none to four), individual child elements `dimensions` have been inserted: instead of metric indicators, the recurrent attribute value `margin` has been associated with another one (`top`, `bottom`, `left` or `right`), in order to locate it on the papyrus sheet[24].

Another papyrological peculiarity is the orientation of the writing in relation to the fibres: along (perfibral) or across them (transfibral). Even if there are some rare excep-

23 For instance:

```
<teiHeader>
<fileDesc>
<sourceDesc>
<msDesc>
<msPart>
<physDesc>
<objectDesc>
<supportDesc>
<support>
<dimensions unit="cm">
<height>11</height>
<width>10</width>
</dimensions>
<rs type="kollesis" n="1">
<measure unit="cm">1.8</measure>
</rs>
</support>
```

24 The situation for a fully preserved papyrus would be the following:

```
<teiHeader>
<fileDesc>
<sourceDesc>
<msDesc>
<physDesc>
<objectDesc>
<layoutDesc>
<layout corresp="#g335-recto">
<dimensions type="margin" n="top"/>
<dimensions type="margin" n="bottom"/>
<dimensions type="margin" n="right"/>
<dimensions type="margin" n="left"/>
</layout>
```

tions[25], the text on a single surface of the vast majority of the documents in the GESHAEM corpus follows the same orientation: so, this characteristic is specified in the XML file part hosting the body of the ancient text, within the `division` which identifies the side of the papyrus[26]. The attribute values `perfibral` and `transfibral` are established in the taxonomy of the corpus[27].

An aspect which stands out, even without knowing the languages involved, is the number of columns and lines: it has been nested in the layout description, on the same hierarchical level as the margins. In the GESHAEM project, it has been assumed that each column of the corpus forms a consistent textual unit: all of them have been represented by a separate element `layout`, have been given an `xml:id`, formed by the XML file name followed by a progressive number, and finally have been described by the number of written lines which they contain and by a plain text in French[28]. Then, in the

25 As an example, the surety contract formed by the fragments Inv. Sorb. 802 + 803b + 1256b + 2733m, which presents a vertical *kollêsis* between two sheets showing different orientations of the fibres: so, the beginning of the lines on the right half of the document is transfibral, while the rest on the left is perfibral.

26 See above, n. 15.

27 In every XML file of the corpus:

```
<teiHeader>
<encodingDesc>
<classDecl>
<taxonomy>
<category xml:id="perfibral">
<catDesc>perfibral</catDesc>
</category>
<category xml:id="transfibral">
<catDesc>transfibral</catDesc>
</category>
```

28 For instance:

```
<teiHeader>
<fileDesc>
<sourceDesc>
<msDesc>
<physDesc>
<objectDesc>
<layoutDesc>
<layout corresp="#g335-recto">
<dimensions type="margin" n="top"/>
<dimensions type="margin" n="bottom"/>
<dimensions type="margin" n="right"/>
<dimensions type="margin" n="left"/>
</layout>
<layout xml:id="g335-1" columns="1" writtenLines="3">
<desc>1 colonne de texte de 3 lignes</desc>
</layout>
```

section devoted to the encoding of the actual text, under the `division` which identifies each side of the papyrus, there are as many child elements as the columns, each one showing a correspondence with the `xml:id` specifically created in the layout description: this is the place where the languages are also declared. The documents from the Jouguet Collection are either in Egyptian demotic or in ancient Greek: the former is indicated with the attribute value `egy-egyd`, the latter as `grc`. Finally, any changing in the hand of the writer is indicated by the element `handShift` immediately preceding the lines of text concerned[29].

Apart from the solutions mentioned above, which pay special attention to the materiality of the papyrus documents, the GESHAEM project follows the Text Encoding Initiative standards and sticks to the EpiDoc guidelines as much as possible[30]. For every fragment, both the modern repository (the Institute of Papyrology of Sorbonne University in Paris[31]) and the discovery site (usually, either Ghôran[32] or Magdôla / Medinet Nehas[33]) are provided: the former inside the child `manuscript identifier` of the element `manuscript part`, together with the inventory and the Trismegistos numbers[34]; the latter

```
<layout xml:id="g335-2" columns="1" writtenLines="32">
<desc>1 colonne de texte de 32 lignes</desc>
</layout>
```

29 As an example:

```
<text>
<body>
<div type="edition">
<div n="1" type="textpart" subtype="fragment">
<div n="r" type="textpart" ana="#perfibral" xml:id="g335-recto">
<div type="textpart" subtype="column" corresp="#g335-1" xml:lang="egy-egyd">
<ab>
<handShift new="m1"/>
<lb n="1"/>
```

30 https://epidoc.stoa.org/gl/latest/.

31 TM Coll 275: https://papyrologie.sorbonne-universite.fr/.

32 TM Geo 715.

33 TM Geo 1284.

34 For instance:

```
<teiHeader>
<fileDesc>
<sourceDesc>
<msDesc>
<msPart>
<msIdentifier>
<country ref="http://geotree.geonames.org/3017382/">France</country>
<settlement ref="http://geotree.geonames.org/2988507/">Paris</settlement>
<institution ref="www.trismegistos.org/collection/275">Paris, Sorbonne, Institut de Papyrologie</institution>
<repository ref="http://thot.philo.ulg.ac.be/concept/thot-4792">Sorbonne
```

nested in the element `history` and its child `origin`[35]. Next to the place of origin the dating is encoded: whenever the fragmentary conditions of the support or the type of text do not allow a precise date, a time span is given using the attributes `notBefore` and `notAfter`. All the texts in the Jouguet Collection come from Ptolemaic necropoleis: as a consequence, their dates are preceded by the minus sign, that is the way of expressing years BC in XML. For the moment, the corresponding Egyptian date is given as plain text within the element `note`[36]. Like most of contemporary humanities computing projects, interoperability is held in due consideration by GESHAEM: whenever possible, links are established with other platforms which ensure stable identifiers for ontological definitions, as THOT[37], or geographical sites, like Trismegistos[38] and Geonames[39].

```
- Institut de Papyrologie</repository>
<idno>0779+1257</idno>
<altIdentifier>
<idno type="TM" corresp="http://www.trismegistos.org/text/4410"/>
</altIdentifier>
</msIdentifier>
```

35 As an example:

```
<teiHeader>
<fileDesc>
<sourceDesc>
<msDesc>
<history>
<origin>
<origPlace cert="low" type="found">
<country ref="http://www.trismegistos.org/place/8881">Egypte</country>
<region ref="http://www.trismegistos.org/place/332">Arsinoite (Fayoum)
</region>
<settlement ref="http://www.trismegistos.org/place/715">Ghoran (Medinet
Ghoran - Kom Medinet Ghuran)</settlement>
</origPlace>
```

36 For instance:

```
<teiHeader>
<fileDesc>
<sourceDesc>
<msDesc>
<history>
<origin>
<origDate notBefore="-0226-02" notAfter="-0226-03">février/mars 226a
<note>Ptolémée III, an 21 = 22, Tybi</note>
</origDate>
```

37 THesauri & OnTology for documenting Ancient Egyptian Resources, hosted by the University of Liège: https://thot.philo.ulg.ac.be.

38 Either Trismegistos Places (https://www.trismegistos.org/geo) or Trismegistos Collections (https://www.trismegistos.org/coll).

39 http://geotree.geonames.org.

In order to provide the entire corpus with consistent cross references, whenever the same entity appears multiple times in different texts, three external authority lists have been compiled up to now: toponyms, persons and legal clauses. Toponyms are univocally identified by an internal `xml:id` and a stable Uniform Resource Identifier (URI) in Trismegistos Places: their different names in demotic, Greek and French are recorded, thus allowing comprehensive researches[40]. Similarly to places, every person is given an `xml:id`, a link to Trismegistos People (if available) and the multiple variants of his name in the three languages previously mentioned: correspondences between demotic and Greek are often supplied by bilingual papyri, like surety contracts. When the context reveals details about a person, the element `roleName` allows one to encode them: the attribute value `activity` is used for juridical relationships like debtor, guarantor or witness; `occupation` for actual professions; `title` for official ranks. For family bonds, it might employ either the attribute value `filiation` for `roleName`, or the separate element `relation` with an attribute value `personal`, then detail the type of relationship (`parent`, `sibling`, `spouse` etc.), the people involved and the hierarchy between them (`active` for ancestors, `passive` for descendants, `mutual` for siblings and spouses)[41]. Finally, in order to compare variants in different legal agreements and to

40 As an example:

```
<text>
<body>
<list>
<item corresp="http://www.trismegistos.org/place/367" xml:id="Athenon">
<label>settlement</label>
<placeName xml:lang="egy-egyd">P3-ʿ.wy-n-Tmtys</placeName>
<placeName xml:lang="grc">Ἀθηνᾶς κώμη</placeName>
<placeName xml:lang="fr">Athenas Kome</placeName>
</item>
```

41 For instance:

```
<text>
<body>
<div type="commentary">
<listPerson>
<person corresp="www.trismegistos.org/person/16258" sex="1" xml:id=
"Nechtenebis58">
<persName>
<name xml:lang="egy-egyd">Nḫt-nb=f</name>
<name xml:lang="grc">Νέχθνιβιν</name>
<name xml:lang="fr">Nechtenebis fils de Pasis</name>
<roleName type="titre">comogrammate</roleName>
<roleName type="occupation">scribe de village</roleName>
<roleName type="activity">garant en second</roleName>
</persName>
</person>
<person corresp="www.trismegistos.org/person/76793" sex="1" xml:id= "Pasis93">
```

make synoptic studies easier, an original authority list has been compiled, entirely focused on clauses: an `xml:id` and a definition have been assigned to each item, inserting mutual exclusions whenever juridically motivated[42]. All these definitions are called back in the XML files through a series of tags `reference`, put around the name or the phrase appearing in the texts.

Like other papyrological portals, the single XML files and the images will be freely accessible under a Creative Commons Licence[43], and the interface available on the website geshaem.huma-num.fr will be user-friendly. GESHAEM aims at giving the edition of some of the texts selected for their historical issues, mainly the surety contracts. The corpus will be available in open-access and accompanied by a paper version: it will not include all the papyri kept by the Institute of Papyrology of Sorbonne University, but will be limited to the documents studied by the project.

One might search for papyri either by inventory or by publication number. Each record will present a brief description in French of the support, of the text, of its provenance and date, then two lists: one for the places mentioned, `Lieux`, the other for the individuals, `Personnes`. Clicking on a button, it will be possible to highlight the former (GEO) or the latter (PER) in the document, displayed on two side-by-side columns: on the left, depending on the language, there will be either a transliteration of demotic or a transcription of Greek; on the right, their translation in French. After the critical apparatus and the commentary, referring to specific lines of the text, both colour and infrared photos will be available and might be magnified on screen. On the bottom left corner of the webpage, there will be the link to Trismegistos Texts, a button which will give access to the XML source file, and the stable URI for every papyrus.

```
<persName>
<name xml:lang="egy-egyd">Pa-sy</name>
<name xml:lang="grc">Πάσιτος</name>
<name xml:lang="fr">Pasis</name>
<roleName type="filiation">père</roleName>
</persName>
</person>
<relation type="personal" active="#Pasis93" name="parent" passive=
"#Nechtenebis58"/>
```

42 A clause of “obligation of payment”, for example, precludes the presence of an “obligation of appearance”, given the different nature of surety contracts bearing them:

```
<text>
<body>
<list>
<item xml:id="OP" exclude="#OA">
<term xml:lang="en">Obligation of payment</term>
<term xml:lang="fr">Obligation de paiement</term>
</item>
```

The same applies to the “penalty for failure to pay” and the “penalty for failure to appear”.

43 https://creativecommons.org/licenses/by/4.0.

For certain types of texts such as accounts, the possibility of automatic calculations using encoding is being explored. In particular, that which has been developed for an online edition of the temple accounts of the village of Soknopaiou Nêsos / Dime during the Roman period, which has been another digital project led by Marie-Pierre Chaufray and Nathalie Prévôt, namely DimeData[44], might be utilised.

Bibliography

Bülow-Jacobsen, A. (2009), *Writing Materials in the Ancient World*, in *The Oxford Handbook of Papyrology*, ed. by R.S. Bagnall, Oxford – New York, 3–29.

Cadell, H. (1966), *Papyrus de la Sorbonne (P. Sorb. I), nos 1 à 68*, Paris.

Cadell, H. – Clarysse, W. – Robic, K. (2011), *Papyrus de la Sorbonne (P. Sorb. III nos 70–144)*, Paris.

Chaufray, M.-P. – Wackenier, S. (2016), *Papyrus de la Sorbonne (P.Sorb. IV No 145–160)*, Paris.

Clarysse, W. – Thompson, D.J. (2006), *Counting the People in Hellenistic Egypt*, Cambridge.

De Cenival, F. (1973), *Cautionnements démotiques du début de l'époque ptolémaïque (P. Dém. Lille 34 à 96)*, Paris.

De Cenival, F. (1984), *Papyrus démotiques de Lille (III), nos 99–108*, Cairo.

Guéraud, O., (1931), *ΕΝΤΕΥΞΕΙΣ. Requêtes et plaintes adressées au Roi d'Égypte au IIIe siècle avant J.-C.*, Cairo.

Jouguet, P. (1901), *Fouilles du Fayoum : rapport sur les fouilles de Médinet-Mâ'di et Médinet-Ghôran*, Bulletin de Correspondance Hellénique 25, 380–411.

Jouguet, P. (1902), *Rapport sur deux missions au Fayôum*, Comptes-rendus des Séances de l'Académie des Inscriptions et Belles-Lettres 46, 346–59.

Jouguet, P. (1907–28), *Papyrus grecs, tome premier*, Paris.

Jouguet, P. – Lefebvre, G. (1902), *Papyrus de Magdôla*, Bulletin de Correspondance Hellénique 26, 95–128.

Lesquier, J. (1912), *Papyrus de Magdôla réédités d'après les originaux*, Paris.

Pirrone, A. (2022), *Apprentissage profond auto-supervisé de métriques : application à la prédiction d'assemblage de fragments de papyrus, thèse de doctorat inédite*, Bordeaux.

Sottas, H. (1921), *Papyrus démotiques de Lille, tome Ier, nos 1–33*, Paris.

Uggetti, L. (2022a), *La collezione dell'Istituto di Papirologia dell'Università della Sorbona e il progetto GESHAEM*, in *Atti del XIX Convegno di Egittologia e Papirologia, Siracusa, 1–4 ottobre 2020*, a c. di A. Di Natale – C. Basile, Syracuse, 321–32.

Uggetti, L. (2022b), *An Unpublished Petition from the Sorbonne Collection*, in *Proceedings of the 29th International Congress of Papyrology, Lecce, 28th July – 3rd August 2019*, ed. by M. Capasso – P. Davoli – N. Pellé, Lecce, II, 989–1001.

44 https://dimedata.huma-num.fr/.

Angelo Mario Del Grosso — Simone Zenzaro — Federico Boschetti — Graziano Ranocchia

Bridging Traditional and Digital Papyrology with Domain-Specific Languages

The GreekSchools Case Study

1 Introduction

Papyrology, the scholarly discipline concerned with the study of ancient texts inscribed on papyri, holds immense significance in reconstructing historical, literary, philosophical, and linguistic aspects of the past. Among the objects of particular interest are the Herculaneum papyri, an extraordinary collection of about 1000 carbonized scrolls containing invaluable Greek philosophical texts.[1]

Modern papyrology has begun to leverage computational capabilities and the well-known Papyri.info platform is the state-of-the-art realization of this direction.[2] While the potential for digital and computational advancements to enhance the workflow of papyrologists is undeniable, their widespread adoption remains somewhat hindered or confined to specific tasks such as the creation of textual archives or printed editions.[3] Presently, digital papyrology relies heavily on shared XML vocabularies, such as TEI/EpiDoc,[4] which often necessitate technical training. Moreover, EpiDoc and related methodologies, like the Leiden+ conventions,[5] diverge significantly from traditional editorial practices.[6]

In this chapter, we propose a method to bridge the divide between traditional and digital papyrology by harnessing the capabilities of Domain Specific Languages (DSLs).[7] Our approach, namely *DSL-based Digital Scholarly Editing* (DSL-based DSE), seeks to pave the way for harmonious integration.[8] We believe that it is possible to bridge the gap between traditional and digital papyrology leveraging Domain Specific Languages by following the DSL-based DSE methodology. Throughout this chapter, we describe the GreekSchools project that is our testing ground for our methodology. Additionally, we offer a succinct theoretical foundation for our novel approach, elucidating the underly-

1 Sider 2005.
2 Berkes 2018, 75–86.
3 Reggiani 2018.
4 https://epidoc.stoa.org. All hyperlinks last accessed on 30.6.2024.
5 https://papyri.info/docs/leiden_plus.
6 Zenzaro 2022.
7 Mugelli – Boschetti – Del Gratta *et al.* 2016, 103–20.
8 Zenzaro – Del Grosso – Boschetti – Ranocchia 2023, 230–2.

https://doi.org/10.1515/9783111070162-009

ing principles guiding its design. Then we define DSL-based DSE and illustrate its practical implications through concrete examples, showcasing the dynamic interplay between our methodology and EpiDoc. Finally, we introduce *CoPhi Editor*, a collaborative and cooperative Web-based platform that implements the DSL-based DSE methodology for the GreekSchools project, but it aims to position itself among the useful tools for collaborative editing of digital scholarly editions, like SoSOL,[9] Perseids,[10] TextualCommunities,[11] and others.[12]

2 The GreekSchools project

The Project ERC Advanced Grant 885222-GreekSchools, The Greek Philosophical Schools according to Europe's Earliest 'History of Philosophy': Towards a New Pioneering Critical Edition of Philodemus' Arrangement of the Philosophers (European Commission, Horizon 2020, Excellent Science, PI: G. Ranocchia, https://greekschools.eu) aims to provide a new critical edition, with introduction and commentary, of Philodemus of Gadara's Arrangement of the Philosophers, a treatise in several books which represents the earliest 'history of philosophy' to have reached us directly from antiquity.[13]

Notoriously, a new critical text requires a long time before stabilizing itself and imposing itself as normative among scholars. In addition, philologists and papyrologists are often accustomed to working alone. In order to overcome these issues, we intend to engage the scholarly community in an ongoing collaborative review process for critical editions by launching CoPhi Editor,[14] a new ad-hoc open-source scholarly Web platform on which the new critical texts will be uploaded and be made available open-access, for papyrological and exegetical comments. The platform allows to collaboratively work, for each section of Philodemus' *Arrangement of the Philosophers*, on a text reconstruction with a critical apparatus and a modern translation and, besides, the corresponding infrared photographs, the new Shortwave-Infrared hyperspectral images, the Oxonian and the Neapolitan apographs, the transcriptions of the 'interpreti' and the print proofs for the Neapolitan *Collectiones* (whenever extant), with interlinking possibilities. This material will be made accessible through permanently open working sessions during which the editions will be revised and improved in a collaborative way and upon registration by all concerned scholars (mostly papyrologists, classical philologists and historians of ancient

9 Baumann 2013.
10 Almas 2017.
11 Robinson – Bordalejo 2016.
12 Del Grosso – Boschetti – Zenzaro – Ranocchia 2023.
13 Ranocchia – Puglia – Vassallo *et al.* 2022.
14 Zenzaro – Del Grosso – Boschetti – Ranocchia 2022, 20–5.

philosophy).[15] The goal is to create a flexible collection of critical texts constantly monitored by the scholarly community by making both electronic texts and digital sources remotely accessible through a single interface with advanced search capabilities.[16] The digital collection will describe all the significant data towards the publishing of a full-fledged scholarly edition of Philodemus' *Arrangement of the Philosophers*.[17]

3 A Hjelmslevian view on papyrology

The study of papyri is inherently interdisciplinary because it requires the joint analysis of both the matter-substance-form of expression and the form-substance-matter of content. The reasons for this chiasmus will be seen throughout this section. The terminology adopted here is clearly of structuralist derivation, according to the definitions of Hjelmslev's *Prolegomena*[18] and the clarifications of Hjelmslev's *Stratification*.[19] Indeed, to the double articulation of F. de Saussure[20] into words (signifiers) and concepts (meanings), Hjelmslev adds the fundamental distinction between form and substance that affects both the plane of expression (where signifiers are constructed) and the plane of content (where meanings are delineated). Hjelmslev mentions – but delegates to disciplines other than linguistics (and semiotics more generally) – the study of matter without form since only substance, that is, matter that has received form, is semiotically relevant. However, the matter of expression is pertinent for the study of the document even when it is illegible, and the matter of content is pertinent to the study of thought even when it is not expressed linguistically.

The primary path (Fig. 1) suggested by Hjelmslev[21] goes up from the substance to the form of expression and then down from the form to the substance of content. The secondary path is backward. Indeed, the first path describes a process of decoding the communicative intention of the original author, which involves an activity of abstraction (to relieve the form from the noise of matter, whether acoustic or scribal) followed by analytical explanation (to associate words with concepts, that is, forms of expression with forms of content) and completed by interpretive synthesis (to embody the abstract concepts, identified thanks to the previous stage, in the material and spiritual reality of the author).

15 Zenzaro et al. 2023, 230-32.

16 Del Grosso et al. 2018, 214-19.

17 The project description, objectives, funding, coordinator, hosting institution (University of Pisa), and co-beneficiaries (National Research Council of Italy and Italian Ministry of Culture) are available on the EU webpage dedicated to funded projects. Specifically, the 885222-GreekSchools project is described at the following web PID: https://doi.org/10.3030/885222.

18 Hjelmslev 1969.

19 Hjelmslev 1954.

20 de Saussure 1995, 97–103.

21 Hjelmslev 1954.

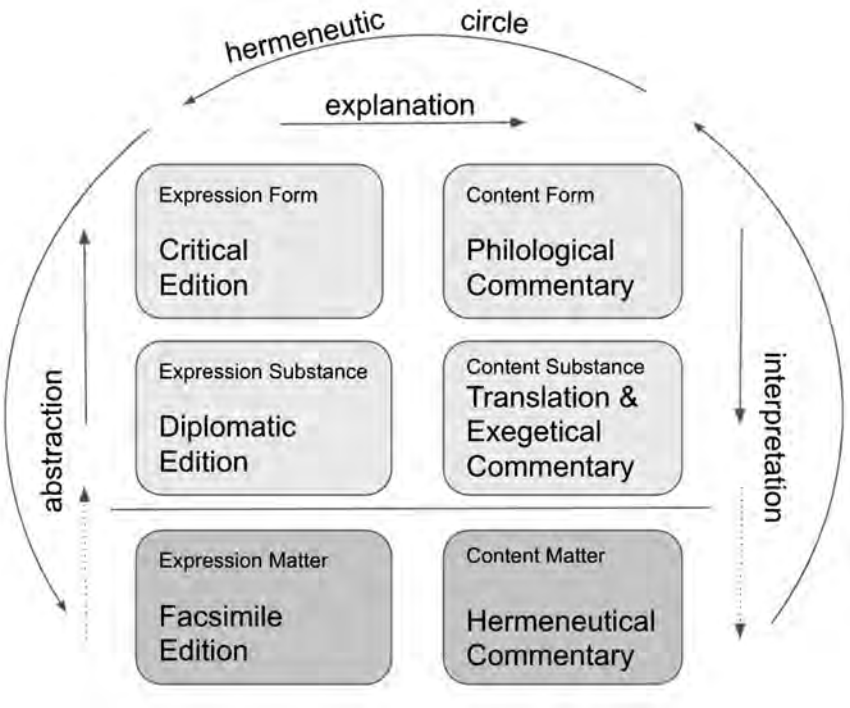


Fig. 1: Hjelmslevian framework for *constitutio textus* and *interpretatio.*

Papyrologists not only add, as mentioned above, an initial stage for the study of the materiality of the medium and a final stage for the understanding of the material and spiritual universe of the author, but they face the path over and over again in both directions (like also philologists do), because a datum acquired on the level of expression modifies (and sometimes distorts) the interpretation and a datum acquired on the level of content helps (and sometimes revolutionizes) the reading. Each phase uses its own methods and tools and produces digital resources available to all scholars involved in the other phases of papyrological studies.

The matter of the expression from the papyrologist's point of view is the papyrus scroll, the ink, and the chemical and physical agents that corrupted the writing medium. Writing techniques and information about preservation contexts also pertain to the matter of expression. The papyrologist must work closely with specialists in scientific disciplines to enhance the legibility of the papyrus by increasing the contrast between the support (background) and traces of writing (foreground). Acquisitions of very high-resolution images (including investigations by microscope), spectrographic images, and three-dimensional models of the scroll are aimed in this direction. Similarly to drawings or pictures in the past, the digital representations reduce the need to continually re-examine the primary sources, although they cannot, obviously, and should never elimi-

nate it altogether. On images that digitally represent the matter of expression, continuous functions can be applied (though in a digital domain they are approximate by discrete functions): increasing or decreasing brightness, contrast, saturation, identifying the contour of a trace, zooming in or out on a region of interest.

The substance of the expression is where the written form shapes the scriptural matter, which in turn resists (confuses, obscures) the form. Such substance is the main object of the diplomatic edition, where allographs, empty gaps and dotted letters are represented. For the diplomatic edition, papyrologists rely on the expertise of palaeographers (or their own notions of palaeography). If the matter of expression is a formless continuum, the substance, on the other hand, as formed matter, is discretizable, although the signs used to describe it do not have oppositional value for the purposes of semiosis (identification of a meaning by a signifier) but do have distinctive value for the study of palaeographical and philological phenomena. The palaeographical apparatus, in its most recent evolution,[22] describes writing traces in a rigorous and unambiguous manner, discretizing the graphic continuum into horizontal, vertical, oblique strokes, etc. within the sextants (anterior, posterior, high, middle and low sectors) in which the alphabetic sign to be reconstructed is inscribed.

The form of the expression finds adequate representation in the critical edition. Indeed, the critical edition is the place where signs have accomplished linguistic value, within the text or the apparatus. Among all the reading possibilities offered by diplomatic editions, the critical edition narrows the field to those readings most likely to make sense. The uncertainty of the dotted letters is resolved and the gaps are filled with conjectural additions. The critical edition concerns itself with the text as an abstract object, that is, as an object, unencumbered by the materiality of the documents.

The form of the content is explained in the philological commentary that accompanies the critical edition. The philological commentary collects the result of morpho-syntactic, semantic, rhetorical and stylistic analyses conducted on the text and relevant to justify the choices made by the editor for the *constitutio* of his edition. In the philological commentary, loci similes (not necessarily expressed identically, but close or identical in content) are given when requested to support a reading in order to clarify its meaning and appropriateness to the context.

The substance of the content is made explicit through scientific translation (total translation, according to Torop)[23] and forms the basis of exegetical commentary. Scientific translation forces the editors to express not only what they have formally understood about the text, but also what they have deeply comprehended, in order to illustrate to a contemporary audience the ancient concepts. An exegetical commentary should take into account all the previous ancient commentaries and modern studies (when available) aimed to the interpretation of the text.

22 Ranocchia 2023.

23 Torop 2010.

Content matter is the subject of hermeneutic commentary, which is aimed at shedding light on the text in question as a product of a broader historical, anthropological, literary and philosophical context. Content matter is a continuum only partially substantiated (and substantiable) in words. Thus, the hermeneutic commentary should take into account data and interpretations provided by disciplines such as archaeology, art history, and historical anthropology, which seek to reconstruct the extralinguistic context and cultural climate of the ancient world in which texts originated.

The necessity to understand the parts (single readings) through the synthetic knowledge of the whole (text in its context) and the whole through the analytic study of its parts (the well-known hermeneutic circle) requires a continuous process of revision (frequent, before publication) or re-edition (sporadic, after publication) of each scientific product (facsimile, diplomatic, critical edition and philological, exegetical, hermeneutic commentary) in light of the new findings (or re-thinkings) contained in all the other.

4 DSL-based DSE

The editing environment conceived for the GreekSchools project requirements is a computational philology platform able to support the traditional editorial process by automatically handling the ecdotic conventions that are already standard for textual scholars.[24]

In this way, the effort necessary to produce the digital representation of the reconstructed text depends mainly on the very nature of the philological process, actually supported and made more effective by the digital medium.[25]

As a consequence, in order to create an edition of a text fragment, scholars, as usual, transcribe the text (*diplomatic transcription*) while describing relevant facts related to the surviving characters visible on the writing surface of the primary source (*palaeographical apparatus*). Afterward, scholars produce a literary transcription and provide authoritative conjectures on missing or damaged characters that are compatible with the available space and vestiges (*philological apparatus*). Finally, the editor may translate the reconstructed text into any modern language.

During each step, scholars produce edited texts containing special conventions identified by standard characters whose meaning is shared within the same research community. These conventions can represent various textual phenomena, such as gaps, uncertain character readings, interlinear additions, readings from apographs, partially readable characters, scribal deletions, and substitutions. Moreover, in the context of philological reconstruction, they may involve characters with alternative readings, text

24 Boschetti – Del Grosso 2020, 65–99.

25 Zenzaro 2022, 20–5.

supplements, editorial substitutions, editorial expunctions, editorial corrections, and more. Over time, additional editorial conventions have been formalized to describe both the palaeographical apparatus and the philological one.

Within the scholarly environment that we are developing as part of the GreekSchools project, editorial phenomena are recognized and handled automatically once they are written in the edition. We named our approach DSL-Based DSE,[26] which consists of the following steps: 1) editing the text using traditional conventions (disambiguation of some conventions may be necessary); 2) processing the text automatically, leveraging domain-specific languages; 3) transforming the recognized text into a custom data format, usually XML; 4) converting the custom representation into TEI/EpiDoc or any other output format, such as Microsoft Word documents (DOCX) or Adobe Portable Document Format (PDF), using the mapping capability of the XSL-T transformation language.

The examples that follow illustrate the DSL-based DSE approach we have conceived to support textual scholars in their daily editorial work. Specifically, we demonstrate the application of this approach within the context of column 64 of P.Herc. 1004, which is being edited by Graziano Ranocchia and Christian Vassallo and visually represented in Figure 2, for the diplomatic transcription and the palaeographical apparatus, and in Figure 14, for the literary transcription and the philological apparatus.

On the right side of Figure 2, we can find the diplomatic transcription of the Greek text extracted from the primary source. On the left side, the palaeographical apparatus is presented. Both pieces of text have been compiled by the editors.

This extract from the "print" edition clearly shows the use of specific editorial conventions to record textual phenomena and editorial facts. For instance, in line 1 of the diplomatic transcription, a lacuna with uncertain length is denoted using a combination of isolated sublinear dots, small round brackets and square brackets. Line 5, instead, presents the recording of uncertain characters (denoted by using isolated sublinear dots), apograph readings (denoted by half square brackets), interlinear insertion (denoted by raised omission brackets). Also worth mentioning here is the presence of a nested apograph reading within the interlinear insertion. Line 6 has a scribal deletion denoted by means of double brackets. Line 8 shows a scribal substitution denoted by juxtaposed deletion and insertion (⟦ ⟧ʽ ʼ). Additionally, between line 11 and line 12 the diplomatic transcription presents an interlinear *diple obelismene*, a diacritical sign sometimes found in papyri.

26 DSL stands for Domain Specific Language. Whereas DSE stands for Digital Scholarly Editing. Consequently, a DSL-based DSE involves the definition of a formal language and the implementation of special tools, called parsers, to recognize the descriptive piece of text encoded by scholars using ecdotic conventions (Berti 2021; Bucchiarone – Cicchetti – Ciccozzi – Pierantonio 2021; Boschetti – Bambaci – Del Grosso *et al.* 2023)

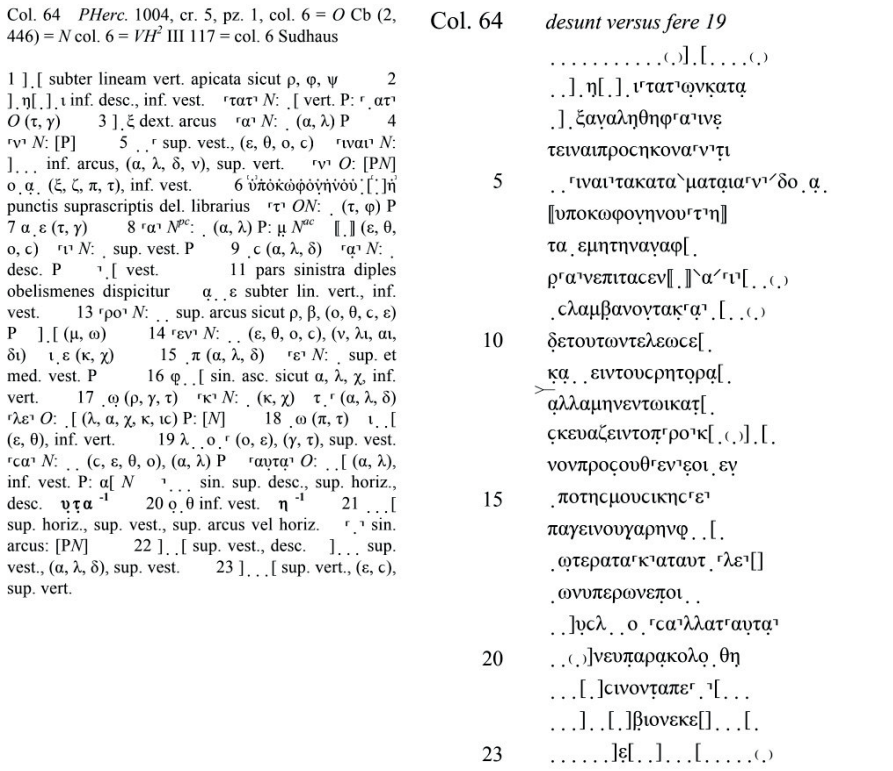


Fig. 2: P.Herc. 1004, col. 64: diplomatic transcription and palegraphic appratus (ed. Ranocchia – Vassallo).

Thanks to the DSL-based DSE approach, scholars can seamlessly employ these conventions as their formal descriptive language to encode the digital text, while also enabling machine actionability.[27] This way, the computational environment provides assistance to the editors during the editing phase. Textual phenomena are automatically recognized by dedicated tools, known as parsers, specifically designed to handle the philological phenomena encoded by the editors.[28]

The first example in Figure 3 represents line 1 of the diplomatic transcription[29] for Column 64 of P.Herc. 1004, as edited by G. Ranocchia and Ch. Vassallo. As mentioned

27 Parr 2014.

28 Zenzaro – Del Grosso – Boschetti – Ranocchia 2022, 20–5.

29 Ideally, the digital edition provides the diplomatic transcription within the `<div type="edition" subtype="diplomatic">` element of the TEI/EpiDoc schema, while the literary transcription is encoded within the sibling `<div type="edition" subtype="literary">` element. Fragments or columns are structurally divided using `<div type="textpart" subtype="fragment">` or `<div type="textpart" subtype="column">` elements.

before, it exhibits a lacuna involving optional ranges, along with the presence of surviving character vestiges.

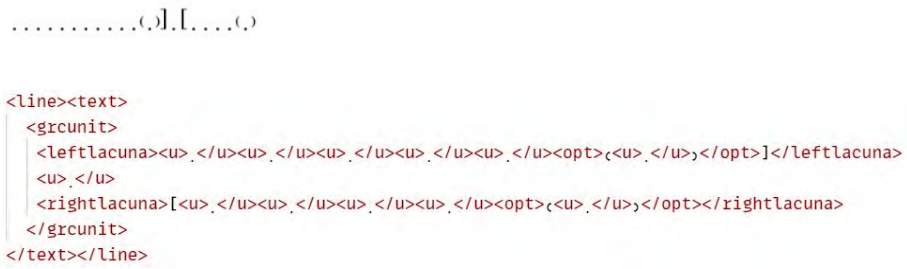


Fig. 3: Line 1 of P.Herc. 1004, col. 64: traditional diplomatic transcription (top) and XML representation exploiting the implemented domain specific language (bottom).

The top section of Figure 3 presents the text as traditionally written for critical editions of Herculaneum papyri, following the conventions used by team members from the GreekSchools project. The row below in Figure 3 displays the XML representation of the automatic recognition process performed by the parser, based on the defined domain specific language.

The XML excerpt shows the tags that explicitly represent the intrinsic data conveyed by the editorial conventions, such as `<line>`, `<leftlacuna>`, `<rightlacuna>`, `<u>`, `<opt>`. These tags are labeled in accordance with the domain experts[30] (in this case, the papyrologists) following the Domain-Driven Design approach that enforces the definition of a shared and ubiquitous language.[31] After parsing the editorial text (which results in an intermediate XML-DSL representation), it undergoes additional processing to generate the TEI/EpiDoc representation of the original edited text.

Figure 4 displays the XML StyleSheet (XSL-T) used to transform the XML-DSL into TEI/EpiDoc format for line 1 of P.Herc. 1004, col. 64. The XSL-T generates an XML fragment that adheres to the formal rules of TEI/EpiDoc specification.[32] In particular, the line is expressed with the element `<lb />`, the lacuna with the element `<gap />` and the attribute `@reason` equals “lost”; the illegible character is represented with the element `<gap />` and the attribute `@reason` equals “illegible”. Additionally, the `@ana`

30 Boschetti – Del Gratta – Del Grosso 2017, 249–53.

31 Domain-Driven Design (DDD) is a software design and development process described by Eric Evans in 2003 (Evans 2003). The process aims to build software applications based on models that represent the concepts and activities of the domain of interest. To achieve this, the approach involves the active participation of domain experts during the application development, with a focus on establishing a shared and unambiguous language called ubiquitous language.

32 Reggiani 2017, 202-54

attribute is used to record the presence of vestiges and the side of the lacuna (left or right). Furthermore, the lacunae are also decorated with the damage tag in order to record the loss of the support.

```
<xsl:template match="line">···
</xsl:template>
<xsl:template match="line/text/grcunit/leftlacuna">
    <xsl:call-template name="edgelacuna" />
</xsl:template>
<xsl:template match="line/text/grcunit/rightlacuna">
    <xsl:call-template name="edgelacuna" />
</xsl:template>
<xsl:template name="edgelacuna">
    <xsl:element name="damage" namespace="http://www.tei-c.org/ns/1.0">
        <xsl:element name="gap" namespace="http://www.tei-c.org/ns/1.0">
            <xsl:variable name="atleast" select="count(child::u)" />
            <xsl:variable name="atmost" select="$atleast + count(opt/u)" />
            <xsl:attribute name="reason">lost</xsl:attribute>
            <xsl:attribute name="extent">
                <xsl:value-of select="$atleast" /><xsl:text> or </xsl:text>
                <xsl:value-of select="$atmost" /><xsl:text> characters </xsl:text>
            </xsl:attribute>
            <xsl:attribute name="atLeast" select="$atleast"></xsl:attribute>
            <xsl:attribute name="atMost" select="$atmost"></xsl:attribute>
            <xsl:attribute name="unit">character</xsl:attribute>
            <xsl:attribute name="ana" select="concat('#',name())"></xsl:attribute>
        </xsl:element>
    </xsl:element>
</xsl:template>
<xsl:template match="line/text/grcunit/u">···
</xsl:template>
```

```
<lb xmlns="http://www.tei-c.org/ns/1.0" n="1"/>
<damage xmlns="http://www.tei-c.org/ns/1.0">
    <gap reason="lost" extent="5 or 6 characters"
    atleast="5" atMost="6" unit="character"
    ana="#leftlacuna" />
</damage>
<gap xmlns="http://www.tei-c.org/ns/1.0"
xml:id="573491" reason="illegible" quantity="1"
unit="character" ana="#vestige" />
<damage xmlns="http://www.tei-c.org/ns/1.0">
    <gap reason="lost" extent="4 or 5 characters"
    atLeast="4" atMost="5" unit="character"
    ana="#rightlacuna" />
</damage>
```

Fig.4: XML Transformation (top) to produce a TEI/EpiDoc fragment of line 1 of P.Herc. 1004, col. 64 (bottom).

In the following examples (Figures 5 to 10) similar XSL-T instructions are used to automatically translate excerpts of the edition into TEI/EpiDoc compliant fragments. The role of XSL-T is analogous to that shown for the example of Figure 4, displaying the entire cycle of transformations performed by the scholarly environment behind the scenes.

Apograph readings, interlinear and unclear characters in line 5 (Figure 5) are automatically recognized by the computational system and encoded using the tags `<apographrdng>`, `<scribins>`, `<uncgrcchar>`, respectively. The corresponding TEI/EpiDoc is presented in Figure 6, where the XML fragment includes the `<gap>` element for encoding the vestiges, the `<supplied>` tag with the `@evidence` attribute equals to "parallel-apograph" and the `<add>` element with a nested `<supplied>` tag. Finally, the TEI/EpiDoc representation includes the `<damage>` tag to indicate certain characters that are incomplete.

All the philological phenomena are treated uniformly. Another example of phenomena in the diplomatic transcription is presented in Figure 7 that shows a scribal deletion, a damaged character, and an apograph reading. It is important to emphasize that the text edited in the platform closely matches what a scholar would have written on paper (or by means of a WYSIWYG[33] word processor), while also being fully machine actionable. The editorial text is recognized by the system and internally represented in XML-DSL format. The output of the automatic conversion to TEI/EpiDoc is shown in Figure 8.

Analogously, for the textual phenomena encoded in line 8, Figure 9 shows, among other phenomena, a deletion and an addition (`<scribdel>` and `<scribins>` tags in XML-DSL) placed close to each other. In this case, the TEI/EpiDoc XML will record a substitution by means of the `<subst>` element together with its children `<del>` and `<add>` elements.

Figure 10 shows the *diple obelismene*, which is encoded in TEI/EpiDoc using a `<milestone>` element with the value of the `@rend` attribute set to "diple-obelismene".

Similarly to the diplomatic transcription, the palaeographical apparatus can also be expressed as a Domain Specific Language to capture the rigorous editorial conventions used by Herculaneum philologists for expressing editorial descriptions.

The palaeographical apparatus is a structured piece of scholarly text that documents all the palaeographical descriptions of the Greek vestiges visible upon the papyrus support, including overlying and underlying layers ('sovrapposti' and 'sottoposti'). Modern sources that witness the original text might complement the character description. In the edition of Herculaneum papyri, these modern sources are indicated by sigla: *O* for the Oxonian apographs and *N* for the Neapolitan drawings standing P for the original source.

33 WYSIWYG stands for What You See Is What You Get.

. . ⸢ιναι⸣τακατα⸌ματ̣ια⸢ν⸣´δο .α̣.

```
<line><text>
  <grcunit>
    <u>.</u><u>.</u>
    <apographrdng>⸢<g>ι</g><g>ν</g><g>α</g><g>ι</g>⸣</apographrdng>
    <g>τ</g><g>α</g><g>κ</g><g>α</g><g>τ</g><g>α</g>
    <scribins>⸌<grcunit>
      <g>μ</g><g>α</g><g>τ</g>
      <uncgrcchar><g>α</g>̣</uncgrcchar><g>ι</g><g>α</g>
      <apographrdng>⸢<g>ν</g>⸣</apographrdng>
      </grcunit>´</scribins>
      <g>δ</g><g>ο</g><u>.</u>
      <uncgrcchar><g>α</g>̣</uncgrcchar><u>.</u>
  </grcunit>
</text></line>
```

Fig. 5: Example of apograph text, interlinear addition and incomplete but certain characters in line 5 of P.Herc. 1004, col. 64 (diplomatic transcription).

```
<gap xml:id="c5492" reason="illegible" quantity="1" unit="character" ana="#vestige"/>
<gap xml:id="c5493" reason="illegible" quantity="1" unit="character" ana="#vestige" />
<supplied evidence="parallel-apograph" reason="lost">ιναι</supplied>
τακατα
<add place="interlinear">ματ<damage>α</damage>ια
  <supplied evidence="parallel-apograph" reason="lost">ν</supplied>
</add>
δο
<gap xml:id="c5511" reason="illegible" quantity="1" unit="character" ana="#vestige" />
<damage>α</damage>
<gap xml:id="c5512" reason="illegible" quantity="1" unit="character" ana="#vestige" />
```

Fig. 6: Example of line 5 of P.Herc. 1004, col. 64 rendered in TEI/EpiDoc (diplomatic transcription).

⟦υποκωφοῡηνουʳτʸη⟧

```
<line><text><grcunit>
  <scribdel>⟦<grcunit>
   <g>υ</g><g>π</g><g>ο</g><g>κ</g><g>ω</g><g>φ</g><g>ο</g>
   <uncgrcchar><g>ν</g></uncgrcchar><g>η</g><g>ν</g><g>ο</g><g>υ</g>
   <apographrdng>ʳ<g>τ</g>ʸ</apographrdng>
   <g>η</g></grcunit>⟧
  </scribdel>
</grcunit></text></line>
```

Fig. 7: Example of scribal deletion in line 6 of P.Herc. 1004, col. 64 (diplomatic transcription)

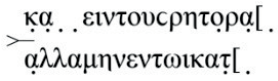
```
<del rend="erasure">υποκωφο<damage>ν</damage>ηνου
  <supplied evidence="parallel-apograph" reason="lost">τ</supplied>η
</del>
```

Fig. 8: Example of the TEI/EpiDoc fragment of the scribal deletion in line 6 of P.Herc. 1004, col. 64 (diplomatic transcription)

ῥʳαʸνεπιτασεν⟦.⟧ˋαˊʳιʸ[..(.)

```
<line><text><grcunit>
  <uncgrcchar><g>ῥ</g></uncgrcchar>
  <apographrdng>ʳ<g>α</g>ʸ</apographrdng>
  <g>ν</g><g>ε</g><g>π</g><g>ι</g><g>τ</g><g>α</g><g>σ</g><g>ε</g><g>ν</g>
  <scribdel>⟦<grcunit><u>.</u></grcunit>⟧</scribdel>
  <scribins>ˋ<grcunit><g>α</g></grcunit>ˊ</scribins>
  <apographrdng>ʳ<g>ι</g>ʸ</apographrdng>
  <rightlacuna>[<u>.</u><u>.</u><opt>(<u>.</u>)</opt></rightlacuna>
</grcunit></text></line>
```

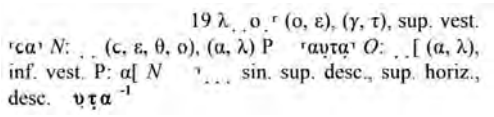
Fig. 9: Example of scribal substitution in Line 8 of P.Herc. 1004, col. 64 (diplomatic transcription).



```
<line>
    <dipleobelismene>⸗</dipleobelismene>
</line>
```

Fig. 10: Example of *diple obelismene* in P.Herc. 1004, col. 64 (diplomatic transcription).

As with the diplomatic transcription, Figures 11 to 13 demonstrate how the phenomena in the palaeographical apparatus are entirely recognized by the domain-specific language that describes the apparatus, translating it into the intermediate XML format (Figure 11) and, subsequently, into TEI/EpiDoc format (Figure 12) through XSL-T instructions encoded within appropriate template rules (Figure 13).



```
<listapp>
  <loc>19</loc>
  <app>
    <lem><lectio>λ..ο.ʳ</lectio></lem>
    <witdetail>
      <palstat>(ο,ε)</palstat>,<palstat>(γ,τ)</palstat>,<palstat><paldesc>sup.</paldesc><paldesc>vest.</paldesc></palstat>
    </witdetail>
  </app>
  <app>
    <lem><lectio>ϛα</lectio></lem><wit>N</wit>:
    <rdg><lectio>..</lectio></rdg><witdetail><palstat>(ϛ,ε,θ,ο)</palstat>,<palstat>(α,λ)</palstat></witdetail><wit>P</wit>
  </app>
  <app>
    <lem><lectio>ʳαυτα</lectio></lem><wit>O</wit>:
    <rdg><lectio>..[</lectio></rdg>
    <witdetail><palstat>(α,λ)</palstat>,<palstat><paldesc>inf.</paldesc><paldesc>vest.</paldesc></palstat></witdetail><wit>P</wit>:
    <rdg><lectio>α[</lectio></rdg><wit>N</wit>
  </app>
  <app>···
  </app>
  <app>
    <lem><lectio>ντα</lectio></lem>
    <witdetail>
      <layer><sub>⁻¹</sub></layer>
    </witdetail>
  </app>
</listapp>
```

Fig. 11: Example of different descriptions recorded within the paleographical apparatus.

```
<listApp>
  <app loc="19">
    <lem>λ..ο.ʳ</lem>
    <noteGrp type="vestigesDesc">
      <note n="1"><choice><unclear>ο</unclear><unclear>ε</unclear></choice></note>
      <note n="2"><choice><unclear>γ</unclear><unclear>τ</unclear></choice></note>
      <note n="3">sup. vest</note>
    </noteGrp>
  </app>
</listApp>
```

```
<app loc="19">
  <lem wit="#N">ϛα</lem>
  <rdg wit="#P">..</rdg>
  <noteGrp  type="vestigesDesc">
    <note n="1"><choice><unclear>ϛ</unclear><unclear>ε</unclear><unclear>θ</unclear><unclear>ο</unclear></choice>
    <note n="2"><choice><unclear>α</unclear><unclear>λ</unclear></choice>
  </noteGrp>
</app>
```

```
<app xmlns="http://www.tei-c.org/ns/1.0" loc="19">
  <lem>υτα</lem>
  <note>underwritten text (⁻¹ stratum)</note>
</app>
```

Fig. 12: Example of a possible TEI/EpiDoc representation of the philological apparatus.

```
<xsl:template match="app">
    <xsl:element name="app" namespace="http://www.tei-c.org/ns/1.0">
        <xsl:attribute name="loc">
            <xsl:value-of select="preceding-sibling::loc"/>
        </xsl:attribute>
        <xsl:apply-templates/>
    </xsl:element>
</xsl:template>

<xsl:template match="lem">
    <xsl:element name="lem" namespace="http://www.tei-c.org/ns/1.0">
        <xsl:value-of select="current()/lectio" />
    </xsl:element>
</xsl:template>

<xsl:template match="witdetail">
 <xsl:choose>
    <xsl:when test="count(child::*) gt 1">···
    </xsl:when>
    <xsl:otherwise>
       <xsl:element name="note" namespace="http://www.tei-c.org/ns/1.0">
          <xsl:apply-templates />
       </xsl:element>
    </xsl:otherwise>
 </xsl:choose>
</xsl:template>
<xsl:template match="layer">
    <xsl:value-of select="if(./sub) then 'underwritten text' else 'overwritten text' " />
    (<xsl:value-of select="concat(sub|sup,' stratum')" />)
</xsl:template>
```

Fig. 13: Example of a possible XSL-T to process palaeographical apparatus.

Col. 64	*desunt versus fere 19*	(*c. 20 linee e 2-3 parole mancanti*) alle
	. . . . . .(.)].[. . . .(.)	cose secondo opinione vera sembra esse-
	. .].η[.].ιτα τῶν κατὰ	re opportuno paragonare ciò che è se-
	δ]όξαν ἀληθῆ φαίνε-	condo vana opinione, ossia ciò che non
	τ' εἶναι προσῆκον ἀντι-	ha relazione con le evidenze, e (sembra)
5	θεῖναι τὰ κατὰ ματαίαν δόξαν,	che i retori non padroneggino assoluta-
7	τά γε μὴ τὴν ἀναφ[ο-	mente nemmeno queste cose. Cionondi-
	ρὰν ἐπὶ τὰς ἐναρ[γεί-	meno, nel trattare la materia, invano
	ας λαμβάνοντα, καὶ [μη-	sembra addurre [analogie prese in presti-
10	δὲ τούτων τελέως ἐ[πι-	to] dall'arte musicale; le cose, infatti,
	κ{α}ρ‹α›τεῖν τοὺς ῥήτορα[c.	non sarebbero state più chiare lasciando
	ἀλλὰ μὴν ἐν τῶι κατ[α-	così come sono (?) quelle intorno alle
	σκευάζειν τὸ προκ[εί]μ[ε-	quali faceva i [suoi] ragionamenti; ma,
	νον πρὸς οὐθὲν ἔοικεν	[da una parte], queste cose [sono per na-
15	ἀπὸ τῆς μουσικῆς ἐ-	tura] facilmente comprensibili a tutti,
	πάγειν· οὐ γὰρ ἦν φαν[ε-	[dall'altra, rispetto] alla vita (*c. 1-2 pa-*
	ρώτερα τὰ κατ' αὐτὰ λε[ί-	*role e 1 linea mancanti*)
	πων, ὑπὲρ ὧν ἐποίει	
	το]ὺς λόγους· ἀλλὰ ταῦτα	
20	μὲ]ν εὐπαρακολούθη-	
	τα π[ᾶ]σιν ὄντα πέφ[υκε,	
	πρὸ]ς δ[὾] βίον εκε[]. . .[.	
23	. . . . .]ε[. .]. . .[. . . . .(.)	

Col. 64 3-8 Sudhaus 6 ⟦ὑπόκωφον ἦν οὐτ' η⟧ del. librarius 7 τά γε Fiorillo: τά τε Armstrong 9-10 [μη]|δέ Sudhaus: [οὐ]|δέ Blank 10-11 ἐ[πι]|κ{α}ρ‹α›τεῖν *: ἐ[ᾶι] | κάπτειν Henry: ἐ[ᾶι] | κρατεῖν Blank, cetera Cirillo 12-16 Cirillo 17-18 λε[ί]|πων *: λέ|[γ]ων Sudhaus 18 ἐποίει *: ἐποιε[ῖ]τ' Sudhaus) 19 Cirillo^{Cpc} 20 Sudhaus 21 πέφ[υκε Sudhaus: πέφ[ηνε Janko, cetera Sudhaus 23 πρὸ]ς δ[έ] *, cetera Sudhaus

Fig. 14: P.Herc. 1004, col. 64: literary transcription, philological apparatus, and translation (edd. Ranocchia – Vassallo).

Literary transcription (Figure 14, left) and the corresponding philological apparatus (Figure 14, bottom) can also be expressed using a domain-specific language. Below, we provide some examples of how it can be managed.

The literary transcription retains some editorial conventions from the diplomatic one, such as the presence of lacunae, or the *diple obelismene*, while introducing or semantically modifying others such as the sublinear dots combined with a Greek character,[34] sublinear asterisk, editorial supplement, addition and deletion.

34 The convention of printing dotted letters within the literary transcription describes an editorial attempt to reconstruct the original text by leveraging on the surviving vestiges while also considering

For example, in line 8, there is an editorial intervention identified by the sublinear asterisk which denotes an apograph reading that has been modified by the editor. Additionally, line 8 includes a supplementation of lacuna and a hyphenation to mark a breaking word at the end of the line (Figures 15–16).

ρὰν ἐπὶ τὰς ἐναρ[γεί-

```
<line>
  <text>
    <grcunit>
      <grcseq><g>ρ</g><g>ὰ</g><g>ν</g></grcseq>
    </grcunit>
    <grcunit>···
    </grcunit>
    <grcunit>
      <grcseq><g>τ</g><g>ὰ</g><g>ς</g></grcseq>
    </grcunit>
    <grcunit>
      <grcseq><g>ἐ</g><g>ν</g><g>α</g></grcseq>
      <editgrcchar><g>ρ</g></editgrcchar>
      <rightsuppl>[<grcseq><g>γ</g><g>ε</g><g>ί</g></grcseq></rightsuppl>
      <hyphen>-</hyphen>
    </grcunit>
  </text>
</line>
```

Fig. 15: Example of editorial correction, editorial supplement and hyphenation (literary transcription).

```
<lb n="8" />ρὰν ἐπὶ τὰς ἑνα<corr resp="#gs-editor">ρ</corr>
<supplied reason="lost">γεί</supplied><lb break="no" rend="-" />
```

Fig. 16: TEI/EpiDoc of editorial correction, editorial supplement and hyphenation (literary transcription).

space constraints. On the contrary, in the diplomatic transcription, dotted letters indicate certain but incomplete readings.

Another interesting example is in line 11, where we encounter an editorial deletion (encoded within curly brackets) and an editorial addition (encoded using angular brackets). Moreover, this line contains uncertain characters represented by dotted letters. At the end of the same line, there is a character supplied by the editor. The representation of the text for this example is shown in Figure 17 (XML-DSL) and Figure 18 (TEI/EpiDoc XML).

One last example (Figure 19) pertains to the philological apparatus and in particular the philological entry at line 7 which encodes two alternative readings, one from Matilde Fiorillo and another from David Armstrong.

κ{α}ῥ‹α›τεῖν τοὺς ῥήτορα[c.

```
<line>
  <text>
    <grcunit>
      <grcseq><g>κ</g></grcseq>
      <editdel>{<grcseq><g>α</g></grcseq>}</editdel>
      <uncgrcchar><g>ρ</g></uncgrcchar>
      <editins>‹<grcseq><g>α</g></grcseq>›</editins>
      <uncgrcchar><g>τ</g></uncgrcchar>
      <grcseq><g>ε</g><g>ῖ</g><g>ν</g></grcseq>
      <grcseq><g>ε</g><g>ῖ</g><g>ν</g></grcseq>
    </grcunit>
    </grcunit>
    <grcunit>
    <grcunit>···
    </grcunit>
    </grcunit>
    <grcunit>
      <grcseq><g>ῥ</g><g>ῥ</g><g>ή</g><g>τ</g><g>ο</g><g>ρ</g><g>α</g></grcseq>
      <rightsuppl>[<grcseq><g>ς</g></grcseq><missingseq><m>.</m></missingseq></rightsuppl>
    </grcunit>
  </text>
</line>
```

Fig. 17: Example of editorial deletion and editorial addition in line 11 of P.Herc.1004, col. 64 (literary transcription).

```
κ<surplus reason="mistake">α</surplus><unclear>ρ</unclear><supplied reason="omitted">α</supplied>
<unclear>τ</unclear>ειν τοὺς ῥήτορα<supplied reason="lost">ς</suppled><gap reason="lost" quantity="1" unit="char">
```

Fig. 18: TEI/EpiDoc of editorial deletion and editorial addition in line 11 of P.Herc. 1004, col. 64 (literary transcription).

7 τά γε Fiorillo: τά τε Armstrong

```
<app loc="7">
  <rdg resp="Fiorillo">τάγε</rdg>
  <rdg resp="Armstrong">τάτε</rdg>
</app>
```

Fig. 19: Example of philological apparatus entry with a lemma and a variant reading (literary transcription).

5 CoPhi Editor

CoPhi Editor is a web-based editing platform designed to enhance the workflow of scholars in creating Digital (and traditional) Scholarly Editions, particularly in the context of the Project ERC Advanced Grant 885222-GreekSchools (https://greekschools.eu). The main objective of CoPhi Editor is to bridge the gap between traditional and digital papyrology implementing the DSL-based DSE approach. To achieve this, it is essential to identify and preserve the valuable aspects of papyrological practices while integrating them seamlessly with the computational capabilities of digital papyrology.[35]

Traditional editing processes primarily focus on the text and its phenomena, which is already a complex task for scholars. An additional layer of encoding in the form of an XML schema can introduce errors and be time-consuming.[36] However, digital papyrology offers several advantages that are worth considering for scholarly editions due to their machine actionability.[37] Tasks such as searching and quantitative analysis become significantly easier with a digital edition compared to a traditional printed one.[38] Moreover, digital editions can be accessed remotely, eliminating the need for physical access to the content of the edition. The digital papyrology community defined shared practices and technologies to formally represent its study objects and encode ancient scholarly editions.[39] They employ the TEI/EpiDoc vocabulary, an XML encoding schema derived from TEI, to represent a vast collection of ancient epigraphical and papyrological sources. This format ensures data interchange, tool compatibility, and long-term preservation. Papyri.info project also adopts the TEI/EpiDoc standard, providing data and metadata from various databases, including DDbDP, HGV, and APIS.[40] The classical pa-

35 Magnani 2018.
36 Boschetti – Del Grosso 2020.
37 Berti 2021.
38 Reggiani 2017, 202–54.
39 Baumann 2013.
40 Baumann 2013.

pyri DCLP section within Papyri.info features authoritative digital editions of the Herculaneum Papyri, with 240 hits for “pherc” documents to date.[41]

When adhering to a standard format like EpiDoc, a digital edition becomes interoperable across various software applications, enabling different types of computations to aid scholars in further studying the edition. It is important to acknowledge that accessing the digital capabilities may require scholars to acquire technical knowledge, such as writing XML files following a schema, and integrating this new knowledge with the traditional approach. However, the risk lies in diverting attention from the text itself towards encoding complexities, potentially compromising the time spent on exploring the textual phenomena.[42]

Between traditional papyrology practices and XML encoding there is a third approach, and CoPhi Editor aims to reconcile both traditional and computational features. The platform preserves the traditional knowledge while incorporating the benefits of computational capabilities. It allows scholars to work in a familiar text-focused environment, while also being machine actionable and interoperable with digital papyrology standards throughout the editing process. CoPhi Editor implements the DSL-based DSE approach. The edition's digital format interoperability is ensured through an automated process that interprets the edition text in a machine actionable format and translates it into the chosen encoding (XML, EpiDoc, Docx, PDF). This approach is similar to that employed in the Proteus project,[43] with further advancements made by utilizing a flexible DSL-based framework. Building on this core workflow, CoPhi Editor provides automated support to enhance the overall quality of the editing process and the final work. It reduces the need for manual checks to ensure consistency in editorial conventions and maximizes the time spent on the most significant textual phenomena of interest.

In the following sections, we provide an overview of the primary functionalities offered by CoPhi Editor.

5.1 Automated check for editorial conventions

The draft proofreading of an edition is an important step before the publication of a critical edition, although it is time consuming. When done thoroughly, the edition is free of typos and all the manual mistakes that eventually come while changing the edited text. Unfortunately, as the amount of text increases, so does the likelihood of accidentally introducing unintended errors.

CoPhi Editor supports this kind of validation of the text, performing automatically a set of validation checks toward the correct form of the edited text. For each error found, a notification to the editor is provided. The types of errors that the platform can

41 Ast – Essler 2018.

42 Boschetti – Bambaci – Del Grosso *et al.* 2023.

43 Williams – Santarsiero – Meccariello *et al.* 2015.

validate depends on the editorial conventions decided by the editor. This way it is possible to eliminate all of the manual mistakes and ensure a consistent application of the editorial conventions. For example, Figure 20 illustrates a syntactic error specifically highlighting a round bracket at the end of the apparatus entry.

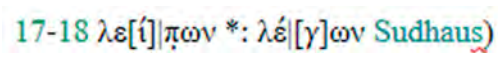


Fig. 20: Example of Critical Apparatus Error.

The platform not only supports single text error checking but also inter-text validation. This kind of validation makes it possible to compare multiple texts against some invariants. For example, an editor can be interested in the correspondence between the presence of an integration in a literary transcription that should derive from the presence of a corresponding missing character. This kind of checks can be performed automatically by the platform.

5.2 Querying data

Before, while and after the editing process, the ability to search information that supports the edition or evaluate further the content is a worthwhile factor. This kind of capability is enabled by a machine actionable representation of the text (e.g. XML).

CoPhi Editor supports querying not only the edited text but also various other sources such as dictionaries, lexica, parallel editions, primary sources, etc. Figure 21 demonstrates the search capabilities utilizing a regex searching mechanism to highlight all the apograph readings within a diplomatic transcription.

5.3 Rich text editor

Making text editing the focus of CoPhi Editor means that the platform should provide a rich text editing experience. And this is the case with CoPhi Editor. Local search and replacement of the text, support for custom fonts (e.g. IFAOGrec for the ancient Greek), highlighting of textual phenomena.

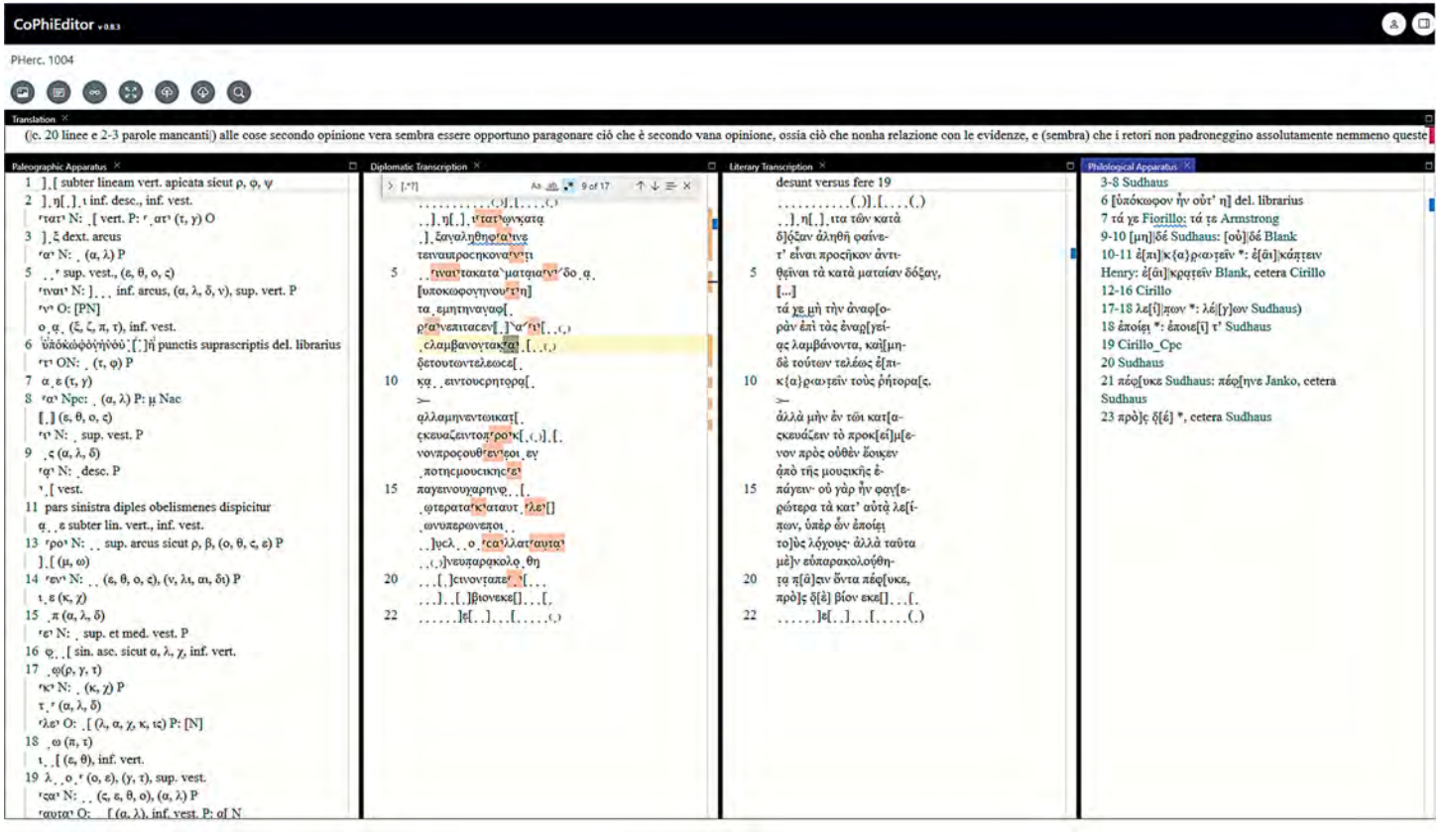


Fig. 21: CoPhi Editor main GUI with the inline search regexp capabilities.

5.4 Cooperation and collaboration

CoPhi Editor supports both cooperation and collaboration.[44] This feature enables remote access to the text bringing the ability to work together on the same edition regardless of the physical distance. Scholars can edit different texts or the same text concurrently. Moreover, a system to open discussion threads gives the scholars the opportunity to bring their respective point of views about the text and trace the contributions. This comment system can be exploited to make an ongoing review process of the text (Figure 22).

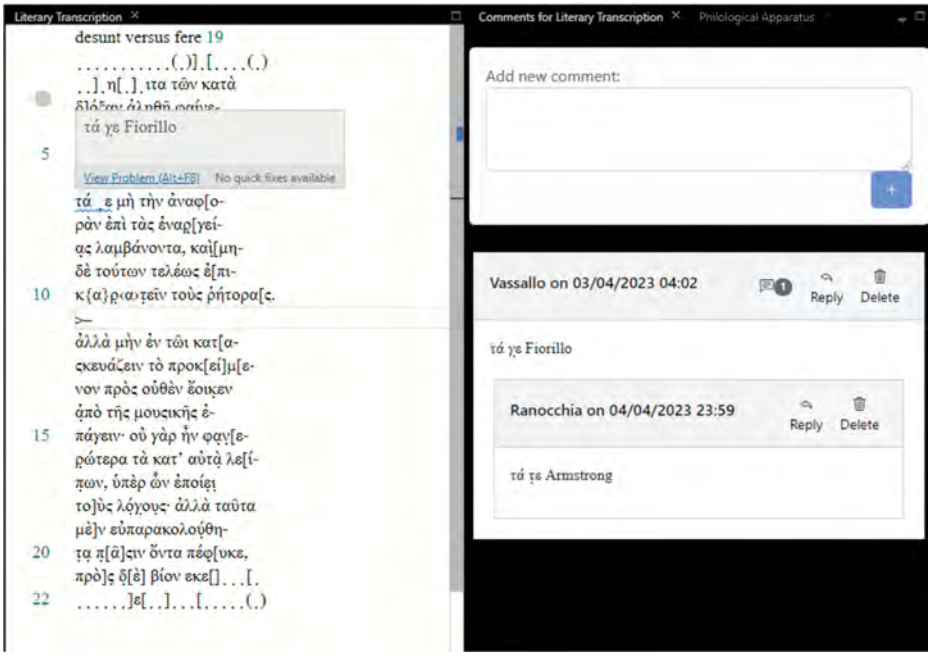


Fig. 22: Comments in CoPhi Editor.

5.5 Dynamic interface layout

Although the graphical user interface (GUI) of the platform in CoPhi Editor tends to be minimalistic, the core editing section is designed to be very dynamic. The goal of such GUI is to permit any arrangement of the text and its sources in the way the scholars see

44 With the term "collaboration", we refer to the involvement of multiple participants in a single task, while with the term "cooperation", we indicate the engagement of multiple participants in multiple subtasks.

fit to work, and possibly mimic, as closely as possible, the physical arrangement counterpart of their traditional workflow. The textual layout can be modified and is personalized for each scholar but a default configuration is proposed based on the observation of the GreekSchools workflow during the textual workshops and seminars held for the project.

The GUI also provides a mechanism to isolate subsets of the texts to reduce the amount of information displayed on screen when needed. For example, in the first stages of a papyrological edition, working only on the diplomatic transcription and the palaeographical apparatus might be preferable.

5.6 Computer assisted editing

The text editor implements a number of text management facilities. At any time, it is possible to open a context menu that shows some proposals to support the human editor. This feature is known as *intellisense*. CoPhi Editor implements a form of intellisense directed to support DSE editing. Via intellisense it is possible to insert special characters (e.g. *diple obelismene*) and to display a number of suggestions to fill in the lacunae.

The suggestions are meant to be an auxiliary tool for the human editor and are based on some metrics and strategies that propose a list of possibilities in context (the position in the text denoted by the cursor). These strategies are based on mutual information such as collocates, distributional semantics, statistics, lexica, n-grams and NLP techniques. The list of suggestions is also based on the DSE text being edited, e.g., the suggestion of which character can be placed in the literary transcription must take into account the palaeographical apparatus where a number of possible options have already been evaluated by the human editor as meeting the palaeographical criteria.

5.7 Agnostic data output

Once the editor decides that the edition is ready to be published, CoPhi Editor supports the publication process by providing a number of output file formats for the edition. This functionality aims to make the edition independent from the publication platform of choice and as a consequence the final edition can comply with the needs of the publisher. If the editor wants to publish the edition in the traditional way, an exporter to DOCX or PDF is the preferred way of transmitting the edition to the publisher. If the editor wants to publish the edition digitally, compliance with EpiDoc or TEI should be possible. CoPhi Editor is able to translate the DSE from its internal data format to each of the abovementioned formats giving the editor both a familiar file format and the most suitable form for the DSE publication.

6 Conclusions

In this chapter, we have presented a comprehensive discussion of the DSL-based methodology, centering our focus on its application within the Project ERC Advanced Grant 885222-GreekSchools. We explored its theoretical underpinnings and provided insights into its current implementation in the CoPhi Editor platform. Through practical exemplification, we highlighted the connections between traditional editorial conventions and the use of TEI/EpiDoc as an automatically generated output in XML format. The CoPhi Editor platform has already hinted its potential to induce a paradigm shift in the existing workflows of papyrologists who are intrigued by the computational capabilities offered in the realm of papyrology but are reluctant to fully leave their familiar editorial practices, which are already standardized within their community. While the platform remains a work in progress, it is poised for further development and enhancement, including the integration of Natural Language Processing capabilities, specifically for *Handwritten Text Recognition* and *Fill Mask* tasks.[45]

Acknowledgments

The ERC Advanced Grant 885222-GreekSchools, *The Greek Philosophical School according to Europe's earliest 'history of philosophy': Towards a new pioneering critical edition of Philodemus' Arrangement of the Philosophers* has received funding from the European Research Council (ERC) under the European Union's Horizon 2020, Excellent Science (Grant agreement No. 885222, PI: G. Ranocchia, https://greekschools.eu).

Bibliography

Almas, B. (2017), *Perseids: Experimenting with Infrastructure for Creating and Sharing Research Data in the Digital Humanities*, Data Science Journal, April 2017, https://doi.org/10.5334/dsj-2017-019.

Ast, R. – Essler, H. (2018), *Anagnosis, Herculaneum, and the Digital Corpus of Literary Papyri*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 63–74. [https://doi.org/doi:10.1515/9783110547450-003]

Baumann, R. (2013), *The 'Son of Suda On-Line'*, Bulletin of the Institute of Classical Studies. Supplement 122, 91–106.

Berkes, L. (2018), *Perspectives and Challenges in Editing Documentary Papyri Online. A Report on Born-Digital Editions through Papyri.Info*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 75–86. [https://doi.org/doi:10.1515/9783110547450-004]

45 Brusuelas 2021.

Berti, M. (2021), *Digital Editions of Historical Fragmentary Texts*, Heidelberg. [https://doi.org/10.11588/propylaeum.898]

Boschetti, F. – Bambaci, L. – Del Grosso, A. M. – Mugelli, G. – Khan, A. F. – Bellandi, A. – Taddei, A. (2023), *Collaborative and Multidisciplinary Annotations of Ancient Texts: The Euporia System*, in *The Ancient World Goes Digital*, 6, Leiden, 172–223. [https://doi.org/10.1163/9789004527119_008]

Boschetti, F. – Del Grosso. A. M. (2020), *L'annotazione di testi storico-letterari al tempo dei social media*, Italica Wratislaviensia (Print) 11, 65–99. [https://doi.org/10.15804/IW.2020.11.1.03]

Boschetti, F. – Del Gratta, R. – Del Grosso, A. M. (2017), *The Role of Digital Scholarly Editors in the Design of Components for Cooperative Philology*, in *Advances in Digital Scholarly Editing*, ed by P. Boot – A. Cappellotto – W. Dillen – F. Fischer – A. Kelly – A. Mertgens – A.-M. Sichani – E. Spadini – D. van Hulle, Leiden, 249–53. [https://www.sidestone.com/books/advances-in-digital-scholarly-editing]

Brusuelas, J. H. (2021), *Scholarly Editing and AI: Machine Predicted Text and Herculaneum Papyri*, Magazén 1, https://doi.org/10.30687/mag/2724-3923/2021/03/002.

Bucchiarone, A. – Cicchetti, A. – Ciccozzi, F. – Pierantonio, A. (2021), eds., *Domain-Specific Languages in Practice with JetBrains MPS*, Berlin. http://www.es.mdh.se/publications/6278-.

de Saussure, F. (1995), *Cours de linguistique générale*, éd. par C. Bally – A. Séchehaye – T. De Mauro, Paris.

Del Grosso, A. M. – Boschetti, F. – Zenzaro, S. – Ranocchia, G. (2023), *GreekSchools: Making Traditional Papyrology Machine Actionable Through Domain-Driven Design*, in *2023 7th IEEE Congress on Information Science and Technology*, Agadir – Essaouira, 621–6. [https://doi.org/10.1109/CiSt56084.2023.10409929]

Del Grosso, A. M. – Bellandi, A. – Giovannetti, E. – Marchi, S. – Nahli, O. (2018), *Scanning Is Just the Beginning: Exploiting Text and Language Technologies to Enhance the Value of Historical Manuscripts*, in *Colloquium in Information Science and Technology, CIST*, New York, 214–9. [https://doi.org/10.1109/CIST.2018.8596373]

Evans, E. (2014), *Domain-Driven Design Reference: Definitions and Pattern Summaries*, Dog Ear Publishing.

Hjelmslev, L. (1969), *Prolegomena to a Theory of Language: Louis Hjelmslev,* transl. by F. J. Whitfield [Rev. English ed.], Madison.

Hjelmslev, L. (1954), *La Stratification du Language*, Word 10, 163–88.

Magnani, M. (2018), *The Other Side of the River Digital Editions of Ancient Greek Texts Involving Papyrus Witnesses*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 87–102. [https://doi.org/doi:10.1515/9783110547450-005]

Mugelli, G. – Boschetti, F. – Del Gratta, R. – Del Grosso, A. M. – Khan, F. – Taddei, A. (2016), *A User-Centered Design to Annotate Ritual Facts in Ancient Greek Tragedies*, Bulletin of the Institute of Classical Studies 59, 103–20. [https://doi.org/10.1111/j.2041-5370.2016.12041.x]

Parr, T. (2014), *Language Implementation Patterns Create Your Own Domain-Specific and General Programming Languages*, Pragmatic Bookshelf.

Ranocchia, G. (2023), ed., *Philosophical Papyri. Journal of Ancient Philosophy and the Papyrological Tradition*, I, Pisa – Rome. http://philosophicalpapyri.libraweb.net.

Ranocchia, G. – Puglia, E. – Vassallo, C. – Pernigotti, C. – Fleischer, K. – Verhasselt, G. – Alessandrelli, M. *et al*. (2022), *The Greek Philosophical Schools According to Europe's Earliest History of Philosophy. Towards a New Pioneering Critical Edition of Philodemus'* Arrangement of the Philosophers, in *XXXth International Congress of Papyrology*, Paris (Poster).

Reggiani, N. (2017), *Digital Papyrology I: Methods, Tools and Trends*, Berlin – Boston. [https://doi.org/doi:10.1515/9783110547474]

Reggiani, N. (2018), *The Corpus of the Greek Medical Papyri and a New Concept of Digital Critical Edition*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 3–62. [https://doi.org/doi:10.1515/9783110547450-002]

Robinson, P. – Bordalejo, B. (2016), *Textual Communities*, in *Digital Humanities Conference*, Kraków, 876–7. http://dh2016.adho.org/abstracts/384.

Sider, D. (2005), *The Library of the Villa dei Papiri at Herculaneum*, Los Angeles.

Torop, P. (2010), *La traduzione totale: tipi di processo traduttivo nella cultura*, transl. by B. Osimo, Milan.

Williams, A. C. – Santarsiero, A. – Meccariello, C. – Verhasselt, G. – Carroll, H. D. – Wallin, J. F. – Obbink, D. – Brusuelas, J. H. (2015), *Proteus: A Platform for Born Digital Critical Editions of Literary and Subliterary Papyri*, in *2015 Digital Heritage*, 2, 453–6. [https://doi.org/10.1109/DigitalHeritage.2015.7419546]

Zenzaro, S. – Del Grosso, A. M. – Boschetti, F. – Ranocchia, G. (2023), *Ease the Collaboration Making Scholarly Editions: The GreekSchools Case Study*, in *La Memoria Digitale: XII Convegno Annuale AIUCD*, Siena, 230–2.

Zenzaro, S. – Del Grosso, A. M. – Boschetti, F. – Ranocchia, G. (2022), *Verso la definizione di criteri per valutare soluzioni di scholarly editing digitale: il caso d'uso GreekSchools*, in *AIUCD 2022 - Proceedings. Culture digitali. Intersezioni: filosofia, arti, media.*, 20–5. [https://doi.org/10.6092/unibo/amsacta/6848]

Riccardo Bongiovanni

A Digital Critical Edition of the Iatromagical Papyri: Opportunities and Challenges

1 Introduction

1.1 The reference corpus

The iatromagical papyri are a subset of the magical texts on papyrus, which is related to the preparation of remedies of magical nature against diseases or physical illnesses.[1] The peculiarity and – in a sense – the difficulty to study such texts lies in the fact that very often the iatromagical texts are not found in papyri of homogeneous content, but in miscellaneous collections, the so-called magical formularies, containing recipes of different nature. On the other hand, they can be found very often among the amulets.[2] However, despite the seeming fragmentation of the material, we are dealing with a proper corpus, provided with very specific and unique characteristics, which differentiate such texts from the wider group of the magical papyri.

First of all, as already said, the purpose of a iatromagical recipe is to heal pathologies of physical nature, and not to face supernatural threats. Second, and because of that, the iatromagical compositions very often utilize a specific vocabulary of scientific origin, with many points in common with contemporary medical writings,[3] not only in the language[4] but also in the layout, which very often recall similar arrangements in the medical or alchemical prescriptions,[5] including the special use of critical signs, abbreviations, and symbols. This makes the editorial project of a uniform corpus reasonable and feasible.[6]

1 See Brashear 1995.

2 See de Haro Sanchez 2004.

3 See Froschauer – Römer 2007; Faraone 2011a/b; de Haro Sanchez 2015a.

4 See de Haro Sanchez 2010.

5 See de Haro Sanchez 2015b; Bongiovanni 2024.

6 Fundamental works for the realization of this project are the critical editions and translations that have been carried out over time on the entire corpus of the Greek magical papyri. Although these works do not delve into the study of iatromagical papyri and the relationship between magic and medicine, they provide the basis for working on texts that have already been corrected and subjected to philological analysis. Unfortunately, most of these works are quite dated. A first comprehensive edition of the Greek magical papyri, published between the late 1800s and the early 1940s, was published by Karl Preisendanz in two volumes titled *Papyri Graecae Magicae*, which were later re-edited by Albert Henrichs (Stuttgart 1972–74). A complete English translation, comprising the Demotic magical papyri as well, was published by Betz 1986. This work was continued by Franco Maltomini and Robert Daniel, who

https://doi.org/10.1515/9783111070162-010

1.2 The project

The project presented here – conducted in the framework of my doctoral fellowship at the University of Pisa under the tutoring of Professor Graziano Ranocchia – is intended to produce a complete and updated digital critical edition of the entire corpus of the iatromagical papyri, including a commentary. In doing this, the aim is to thoroughly analyze each papyrus, collecting every form of relevant data, regarding both the papyrus in its materiality (archaeological context, quality of the writing support and ink, dating) and the text format and its content.[7] Similarly, the palaeographic analysis and the editorial history of each document will not be overlooked.

Besides the obvious commitment to providing a text as accurate as possible, the project aims to analyze the cultural dimension from which these particular texts originate, identifying the influences and antecedents from which such products derive, starting with medical science, magical knowledge, practical procedures, and also literary testimonies. Additionally, the project seeks to investigate the processes that led to the integration of medical-scientific and literary vocabularies within the magical texts, identifying the main sources used by magicians to form their professional language. In doing so, the goal is to contribute to change a stereotypical image of the ancient sorcerers, which is still deeply rooted in portraying them as uneducated or lacking in higher education, and instead restore them to a more accurate dimension, well integrated within the literary and scientific influences of their time.

Overall, it is a multidisciplinary project that intends to encompass philology, papyrology, lexicological and grammatical studies, moving each of these disciplines in synergy with the most modern tools and reflections born within Digital Humanities. The ultimate goal is to create a comprehensive work that can combine the best potentials of traditional philology and the new digital frontiers, thus becoming a useful, updated, and constantly updateable tool, an important reference point in the study of iatromagical papyri, and a solid foundation for further research.

Currently, the project is focused on the digital encoding of the texts on the Papyri.info platform. The choice to prefer a digital edition rather than a traditional printed edition lays in various extra opportunities that a digital editorial environment can offer. A platform like Papyri.info offers several well-known and commonly recognized advantages like the possibility of a continuous update, ease of consultation, and – last but definitely not least – free and open accessibility.[8] Nevertheless, it presents also some

compiled the *Supplementum Magicum*, in two volumes (Köln 1990–92), which includes the magical papyri edited and published separately between 1943 and 1989. A new collective editorial enterprise devoted to the whole Greek and Demotic magical papyri has in the meantime started under the guidance of Christopher A. Faraone and Sofia Torallas Tovar, with currently two volumes, one of general studies and one of text editions (Faraone – Torallas Tovar 2022a/b).

7 On text and context of the iatromagical papyri see de Haro Sanchez 2012.

8 See Reggiani 2017, 222–41.

technical and methodological criticalities, due to the specific history of its origins: being essentially the heir of a database designed for documentary papyri, the encoding of different typologies of texts, with different characteristics and editorial needs, is sometimes challenging.[9]

In this chapter, I will discuss two particular instances raised by the experience of the described project: an opportunity – the possibility to work on more than one interpretive layer – and a challenge – the correct and critical encoding of special magical symbols. This will show how working on special categories of texts raises interesting questions related to the digital representation of the papyri.

2 An opportunity: multi-layer editing

In short words, by combining the scientific accuracy of philology with the functionalities of computer science, the aim is to achieve the abovementioned results by processing the texts on multiple hypertextual annotative and interpretive levels. The first level will present the papyrus itself, with its high-definition image. Subsequent levels will cover the material aspects of the papyrus (palaeography and paratext, layout, and, where possible, the place of discovery); the diplomatic transcription of the papyrus, its critical edition, and finally, a dedicated commentary, divided into four different aspects: philological-literary, medical, and magical-religious, each integrated with grammatical, syntactic, and lexical studies. In the final structure, these layers can be considered together or selectively viewed based on preference.

The critical apparatus of the edition will also not be structured as in a traditional critical edition but will exploit the potential of different hypertextual levels. This approach aims to highlight the editorial history of the texts, presenting, for each papyrus when applicable, the philological reconstructions of various editors, following the model already proposed for the online edition of the Derveni Papyrus sponsored by the Center for Hellenic Studies of the University of Harvard and published on the iMouseion Project platform.[10] As is known, this edition follows a comparative approach, i.e., collecting the various editions and translations of the text in one portal, with the further possibilities to select and display one specific critical edition (Fig. 1), to compare two different editions with each other (Fig. 2), to display a critical edition flanked by its translation (Fig. 3). It is also possible to choose whether to display the text supplied in lacuna or not.

The multiversion structure offers the additional advantage of freeing the editor from the obligation to choose between different reconstructions of a text when such a

9 See Reggiani 2017, 250–4.

10 https://chs.harvard.edu/derveni-papyrus-introduction. See Reggiani 2017, 246–7.

choice is particularly uncertain. In traditional philology, the editor is always 'forced' to choose between a series of text reconstructions, no matter how challenging this task may be, or to leave the text unchanged, marking *cruces* when no reading is sufficiently strong. In cases of significant uncertainty, the use of a multitextual structure allows for a radically different approach: no single variant is imposed over another, nor are *cruces* imposed; instead, all the most plausible reconstructions are presented and justified on the same level, allowing them to be viewed and utilized in subsequent studies.

The main problem is that currently Papyri.info does not support multiversion or multitextual developments. However, other platforms have been designed to work toward that direction. For example, READ (Research Environment for Ancient Documents) is an open-source software, developed by Stefan Baums (Munich), Andrew Glass (Seattle), and Stephen White (Venice), specifically designed for the study of ancient texts on their physical supports.[11] It allows to connect the image of a written artefact with its transcription (Fig. 4), to manage more than one transcriptions of the same object in parallel (Fig. 5), to connect the original texts with translations, and to produce glossaries and palaeographical resources. It also allows for the creation of many annotation layers so to perform different analyses of the texts (palaeography, syntax, vocabulary, etc.).

Fig. 1: iMouseion Project, display of a single edition of the Derveni Papyrus.

11 See https://github.com/readsoftware/read.

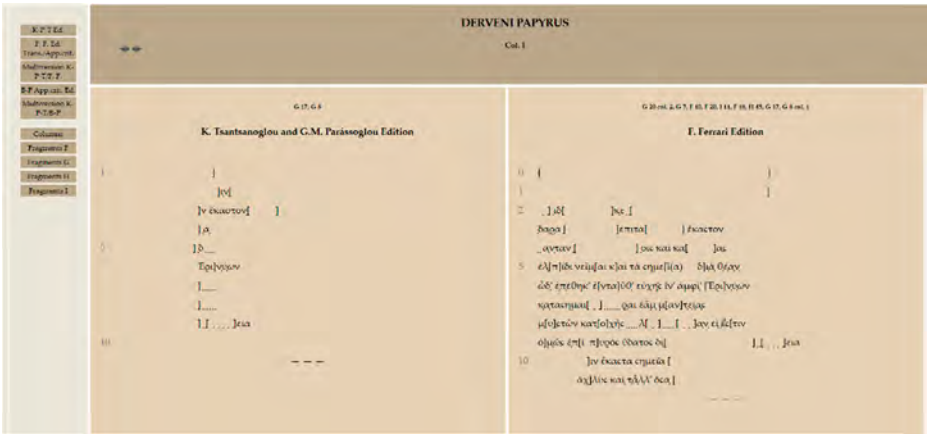


Fig. 2: iMouseion Project, parallel display of two editions of the Derveni Papyrus.

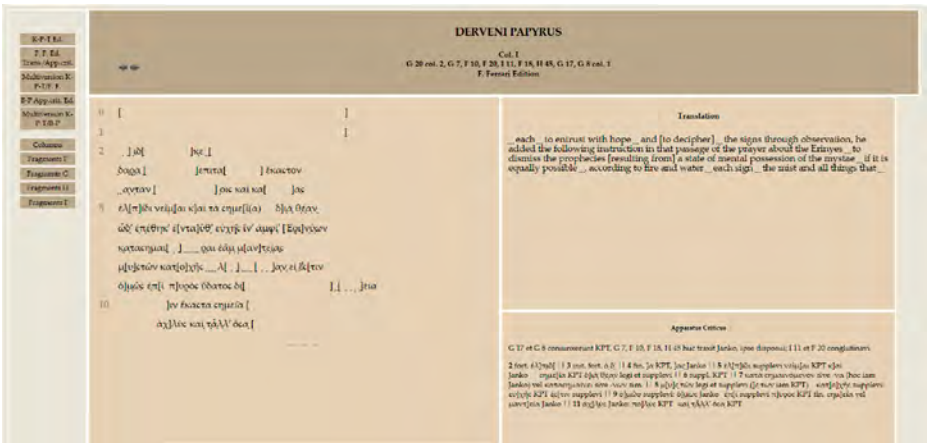


Fig. 3: iMouseion Project, display of an edition of the Derveni Papyrus, with translation and apparatus criticus.

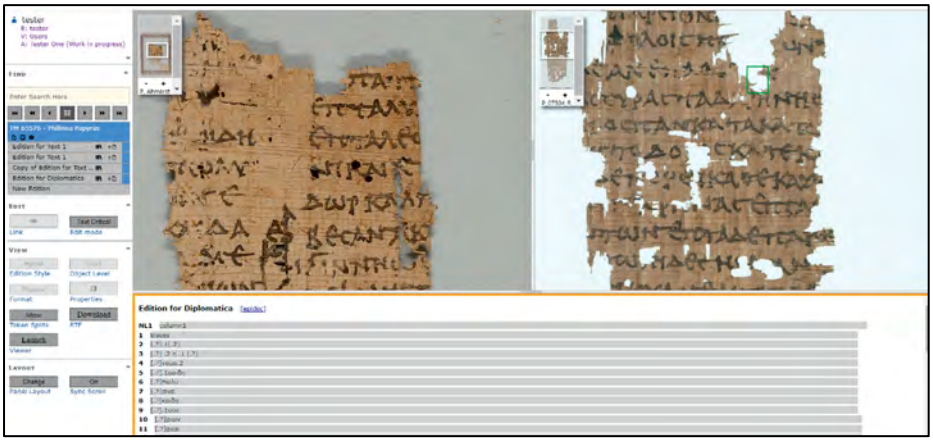


Fig. 4: READ platform, text-image alignment for the Philinna Papyrus (TM 65576).

Fig. 5: READ platform, different edition layers for the Philinna Papyrus (TM 65576).

3 A challenge: the encoding of magical symbols

The magical papyri are often accompanied by drawings and symbols, which explain or enhance their supernatural power. For the sake of this discussion, let us consider the so-called *charakteres*, a series of alphabetical or letter-like symbols – though lacking any semantic or phonetic correlations – that occur very often in the magical texts,[12] in-

12 See Gordon 2014; Frakfurter 2019; Németh 2020.

cluding the Greek magical papyri as a whole, and in some iatromagical text too (Fig. 6). The character forms are mostly nonsensical and may include letters, asterisks, circles, points, closed shapes, strokes, lines. In her 2013 PhD dissertation, Kirsten Dzwiza studied 94 magical texts and recorded 699 different *charakteres* occurring over 943 times.[13]

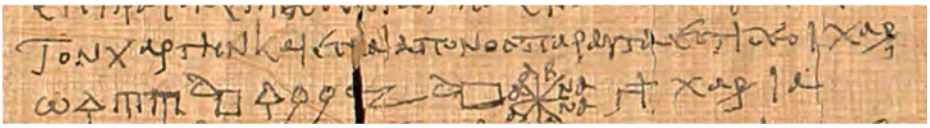


Fig. 6: The long sequence of *charakteres* in a iatromagical prescription against scorpion stings (P.Lond. 121, v 195–196).

Currently, the way of encoding the *charakteres* in Papyri.info is to use the Epidoc XML tag <g> ("glyph"), used to label non-alphabetical and non-standard characters, with an attribute `type` that specifies the name of the symbol.[14] For example, this is used to encode the coronis symbol: `<g type="coronis"/>` in the source XML, `*coronis*` in the Leiden+ markup, which is displayed as ((coronis)) in the Papyrological Navigator output. The current occurrences of *charakteres* in the database are very few. The examples found at the moment are a magical copper *lamella* (TM 285014 = GEMF 26, Akrai, Sicily, 2nd/3rd cent. AD)[15] and a section of a magical formulary on papyrus roll (TM 60204 = P.Lond. I 121 = PGM II 7, Thebes, 3rd/4th cent. AD).[16] In the former (Fig. 7), we find the magical symbols encoded as `<g type="charakteres"/>`; in the latter (Fig. 8), they are differently encoded as generic images, with the markup `<figure><figDesc>charakteres</figDesc></figure>`. Even more generic is a third occurrence of *charakteres*, which cannot be found with a specific query because they are encoded just as "drawings" (TM 63932 = MPER I 28a = PGM II 63 = GEMF 29):[17] `<figure><figDesc>drawings</figDesc></figure>` (Fig. 9).

These are very generic encoding options that do not render the original appearance of the texts, the arrangement of the symbols, their number, and their nature. A user must necessarily refer to a picture of the papyrus or to an autoptic inspection. A different option could be the encoding of each single symbol as a <g> type (Fig. 10). This would allow to keep the exact number of *charakteres* in the text and to have a rough idea of their arrangement, but it is clear that it would not be the optimal rendering.

13 Dwiza 2013.
14 See Reggiani 2017, 252.
15 https://papyri.info/dclp/285014.
16 https://papyri.info/dclp/60204.
17 https://papyri.info/dclp/63932.

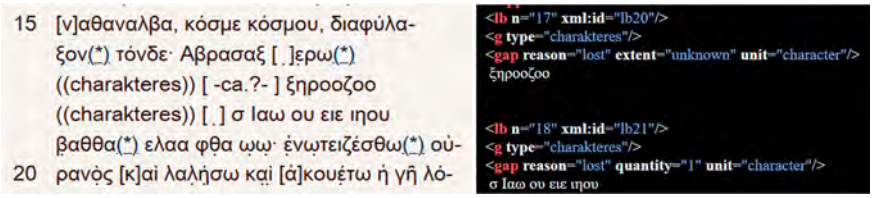


Fig. 7: Display and encoding of *charakteres* as glyphs in https://papyri.info/dclp/285014.

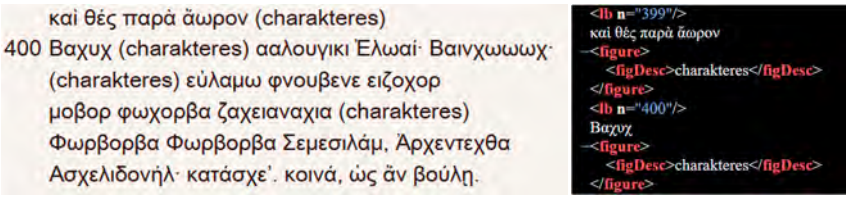


Fig. 8: Display and encoding of *charakteres* as images in https://papyri.info/dclp/60204.

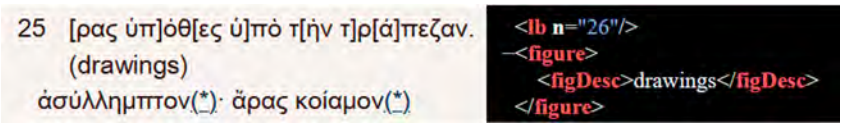


Fig. 9: Display and encoding of *charakteres* as "drawings" in https://papyri.info/dclp/63932.

Fig. 10: Multiple encoding of a sequence of *charakteres* (sample preview display of P.Lond. I 121, see Fig. 6).

In my opinion, the best way of dealing with this issue would be to develop a categorization of the most common *charakteres*, so to including them among the available markup alongside, e.g., the diacritical marks. It could be reached by providing each *charakter* with a standard denomination, which could be used for the `type` attribute of the `<g>` tag. It could also be possible to envision the development of special graphic signs that could display the exact visual rendering of the *charakteres*, as now happens with some diacritical marks (*diple*, *diple obelismene*, etc.).

Bibliography

Betz, H. D. (1986), *The Greek Magical Papyri in Translation, Including the Demotic Spells*, Chicago – London.

Bongiovanni, R. (2024), *Ricettari medici e rimedi magici: un rapporto in evoluzione*, in *Materialità della medicina antica. Aspetti grafici e materiali dei papiri medici dall'antico Egitto*, ed. by N. Reggiani, Parma, 169–88.

Brashear, W. (1995), *The Greek Magical Papyri: An Introduction and Survey; Annotated Bibliography (1928–1994)*, in *Aufstieg und Niedergang der römischen Welt*, ed. by H. Temporini – W. Haase, II, 18.5, Berlin – New York, 3380–684.

de Haro Sanchez, M. (2004), *Catalogue des papyrus iatromagiques grecs*, PapLup 13, 37–60.

de Haro Sanchez, M. (2010), *Le vocabulaire de la pathologie et de la thérapeutique attesté dans les papyrus iatromagiques grecs: l'exemple des fièvres, des traumatismes et de l'« épilepsie »*, BASP 47, 131–53.

de Haro Sanchez, M. (2012), *Mise en contexte des papyrus iatromagiques grecs : recherches sur les conditions matérielles de réalisation des formulaires et des amulettes*, in *Actes du 26e Congrès Internationale de Papyrologie (Genève, 16–21 août 2010)*, ed. by P. Schubert, Geneva, 159–70.

de Haro Sanchez, M. (2015a), *Magie et pharmacopée: l'utilisation des végetaux dans les papyrus iatromagique grecs*, Mythos 9, 149–72.

de Haro Sanchez, M. (2015b), *Between Magic and Medicine: The Iatromagical Formularies and Medical Receptaries on Papyri Compared*, ZPE 195, 179–89.

Dzwiza, K. (2013), *Schriftverwendung in antiker Ritualpraxis anhand der griechischen, demotischen und koptischen Praxisanleitungen des 1. - 7. Jahrhunderts*, PhD Diss., Universität Erfurt.

Faraone, C. A. (2011a), *Magic and Medicine in the Roman Imperial Period: Two Cases of Study*, in *Continuity and Innovation in the Magical Tradition*, ed. by S. Shaked – G. Bohak, Leiden, 135–58.

Faraone, C. A. (2011b), *Magical and Medical Approaches to the Wandering Womb in the Ancient Greek World*, Classical Antiquity 30, 1–32.

Faraone, C. A. – Torallas Tovar, S. (2022a), eds., *The Greco-Egyptian Magical Formularies. Libraries, Books, and Individual Recipes*, Ann Arbor.

Faraone, C. A. – Torallas Tovar, S. (2022b), eds., *Greek and Egyptian Magical Formularies: Text and Translation*, I, Berkeley. [https://escholarship.org/uc/item/9650x69r]

Froschauer, H. – Römer, C. (2007), *Zwischen Magie und Wissenschaft. Ärzte und Heilkunst in den Papyri aus Ägypten*, Vienna.

Gordon, R. (2014), *Charakteres between Antiquity and Renaissance: Transmission and Reinvention*, in *Les savoirs magiques et leur transmission de l'antiquité à la Renaissance*, ed. by V. Dasen – J.-M. Spieser, Florence, 253–300.

Frankfurter, D. (2021), *The Magic of Writing in Mediterranean Antiquity*, in *Guide to the Study of Ancient Magic*, ed. by D. Frankfurter, Leiden – Boston, 626–58.

Németh, G. (2020), *Charaktêres on Curse Tablets in the Western Provinces of the Roman Empire*, in *Choosing Magic: Contexts, Objects, Meanings. The Archaeology of Instrumental Religion in the Latin West*, ed. by R. Gordon – F. Marco Simón – M. Piranomonte, Rome, 125–38.

Reggiani, N. (2017), *Digital Papyrology I. Methods, Tools and Trends*, Berlin – Boston.

Erik Henriksson — Marja Vierros

PapyGreek Search: Exploring the Language of Greek Papyri

1 Introduction

The growing interest in the language of the Greek papyri is closely tied to the increased availability of text corpora and digital research tools. The digitisation and open-access publication of all edited documentary papyri through Papyri.info has laid the groundwork for all subsequent corpus-based linguistic studies.[1] Recently, further progress has been made in the preprocessing of papyrological texts for linguistic analysis,[2] the creation of linguistically annotated corpora,[3] and the development of query tools.[4] Despite these advancements, there is arguably still room for new digital resources.

In this chapter, we introduce PapyGreek Search, a tool designed for papyrologists and linguists interested in exploring the language of Greek documentary papyri. Developed as part of "The Digital Grammar of Greek Documentary Papyri" project,[5] PapyGreek Search is freely accessible online through the PapyGreek project website.[6] PapyGreek Search is distinguished by its search interface that enables simultaneous queries on morphosyntactic constructions, and through the editorial interventions encoded in these texts, phonological and morphosyntactic variation. Additionally, PapyGreek Search provides an interface for visually building syntactic tree queries, which allows users to utilize its treebank search feature without needing to learn a treebank query language.

The chapter is structured as follows. Section 2 outlines the motivation for developing PapyGreek Search, explores related work, and highlights the tool's key functionalities. Section 3 details the technical implementation, including text preprocessing, database architecture, and search algorithms. Section 4 demonstrates the user interface and Section 5 discusses example queries. Section 6 concludes.

1 Evans – Obbink 2010.

2 Vierros – Henriksson 2017; Vierros 2018

3 E.g. Vierros – Henriksson 2021; Keersmaekers – Depauw 2017; Keersmaekers – Van Hal 2023.

4 E.g. Depauw – Stolk 2015; Keersmaekers – Mercelis – Swaelens – Van Hal 2019

5 This project, led by Marja Vierros, received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 758481).

6 https://papygreek.com/search. All hyperlinks last accessed on 21.7.2024, unless differently indicated.

https://doi.org/10.1515/9783111070162-011

2 Previous work and main features

We start by outlining the evolution of query tools for Greek papyri, from the earliest searchable digital databases to contemporary online platforms such as Papyri.info and Trismegistos.org, along with treebank query tools. We then describe the main functionalities of PapyGreek Search.

2.1 Previous query tools for Greek papyri

Digital papyrology started in 1983 with the establishment of the Duke Databank of Documentary Papyri (DDbDP). In the early 1990s, the digitized texts were distributed as CD-ROM copies, searchable using the software available at the time.[7] A key step towards wider access occurred in 1996 when the texts were migrated to the online Perseus Project, which had a basic string search interface.[8] In the 2000s and early 2010s, advances in open-source standards and interoperability led to more sophisticated search capabilities for the papyri. The "Integrating Digital Papyrology" project converted the texts into machine-readable EpiDoc XML format, integrated the Heidelberger Gesamtverzeichnis (HGV) metadata with the DDbDP texts, and eventually created Papyri.info.[9] This platform has become the central hub for editing and searching texts, with its Papyrological Navigator offering a full-text search interface with document metadata filters.[10]

The EpiDoc XML-formatted documentary papyri contain a wealth of information on linguistic variation across domains such as orthography, phonology, and morphosyntax, encoded through various editorial interventions. Depauw – Stolk 2015 introduced the first online tool for systematically querying the orthographic variants, named Trismegistos Text Irregularities (TI).[11] This tool effectively superseded traditional reference books that contain manually collected – and by now partly outdated – lists of non-standard linguistic forms found in the papyri.[12] TI enables users to identify all variant forms encoded in the XML source files by character-level differences between transcriptions and modern corrections (such as αι corrected to ε) and to further refine the results by the context in which they appear. However, a limitation of the tool is that it relies on a relatively old snapshot of the DDbDP data (2016).[13]

7 Van Minnen 1996.
8 Sosin 2010.
9 Baumann 2013
10 https://papyri.info/search.
11 https://www.trismegistos.org/textirregularities.
12 E.g. Mayser – Schmoll 1970; Gignac 1976.
13 https://www.trismegistos.org/textirregularities/methodology.php.

Regarding morphosyntactically parsed documents (i.e. treebanks), a variety of search options is available. General treebank query tools include ANNIS3,[14] PML-Tree Query,[15] TüNDRA,[16] and INESS.[17] Many of these programs allow users to search within their own treebanks, including papyrological ones,[18] provided the data is in a compatible format. Recent additions include the stand-alone DendroSearch tool,[19] specifically designed for Ancient Greek treebanks, and the forthcoming KTB tool.[20] These programs typically employ a custom query language that users must learn to construct queries, which may be challenging for non-technical users. Additionally, these tools require downloading and setup to function properly in the user's software environment.

2.2 Main features of PapyGreek Search

PapyGreek Search complements the previously available toolset for the linguistic study of Greek papyri, offering partly overlapping and partly novel features as compared to other available tools. Its main features are as follows:

- **Papyrological focus:** PapyGreek Search introduces some unique features specifically developed for the linguistic exploration of papyrological texts. Most importantly, its treebank search function includes an integrated text irregularities search, crucial for the linguistic analysis of papyrological texts given their frequent misspellings and other nonstandard word forms. In addition, the search can be further narrowed down by various metadata, such as date, provenance, author and/or scribe.
- **User-friendly interface:** The platform combines complex search features with a simple graphical user interface, aiming to make linguistic analysis of papyrological texts accessible to a wide range of users, including those with limited prior experience with specialized query languages.
- **Advanced features:** Most of the search parameters in PapyGreek Search may optionally be written using regular expressions,[21] which can be used to craft complex queries. Users can also specify word order in the treebank queries, an important and relatively understudied aspect of syntactic structures in Greek.[22]
- **Synchronization:** The PapyGreek database is updated weekly based on changes made to the source texts (available from https://github.com/papyri).

14 Krause – Zeldes 2016.
15 Pajas – Štěpánek 2009.
16 Martens 2013.
17 Rosén *et al.* 2012
18 E.g. Vierros – Henriksson 2021.
19 Keersmaekers – Mercelis – Swaelens – Van Hal 2019.
20 Yordanova forthcoming; Vierros – Yordanova 2022.
21 E.g. Goyvaerts – Levithan 2012.
22 E.g. Vierros – Yordanova 2022.

– **Open source:** PapyGreek Search is fully open source. This not only ensures transparency but also allows for continued development by the academic community. The source code for the PapyGreek system is available at https://github.com/erikhenriksson, and feature requests can be sent to papygreek.helsinki@gmail.com.

3 Technical implementation

This section describes the technical implementation of PapyGreek Search. Those primarily interested in the user interface may skip directly to Section 4, though understanding how the system works may help with constructing queries. The section begins with a description of preprocessing the source DDbDP files for linguistic annotation, followed by the implementations of textual irregularities and treebank search. Finally, the database structure is described.

3.1 Preprocessing: from EpiDoc XML into tokens

The EpiDoc XML schema used by the DDbDP includes sections for document divisions, editorial corrections, gaps, and other pertinent elements in papyrological texts, making it a suitable format for *representing* structured texts. However, XML is not efficient for *searching*; databases are typically better for this purpose.[23] For the PapyGreek platform, we chose MySQL as our database backend due to its quick and reliable handling of complex queries, as well as ease of management. MySQL, a structured database, associates each data point (e.g., a single document or word) with predefined data fields. Thus, we had to first convert the DDbDP source files into structured entities that could be stored in the database.

The first and most crucial step in this preprocessing stage is *tokenization*, meaning the texts' conversion into discrete units, in our case words (or 'tokens') and sentences. The word is a prerequisite level of description for a system where users search for words and their morphosyntactic characteristics, and the sentence level is additionally required for searching treebank annotations, which describe the syntactic relationships of the words within sentences. Typically, word tokens are found by simply splitting the text by white-space. However, Greek has many exceptions to this rule, such as *craseis* or merged words (καὶ ἐγώ → κἀγώ "and I"), which are orthographically one but linguistically two words (also called 'multi-word tokens'). Furthermore, some of the DDbDP

23 The search capabilities of the Papyrological Navigator (https://papyri.info), for example, are also powered by a database (Apache SOLR).

files contain errors and inconsistencies in spacing.[24] Our custom tokenizer addresses these exceptions by rule-based heuristics.[25]

As Vierros – Henriksson 2017 and Vierros 2018 suggest, preprocessing papyrological texts for linguistic analysis benefits from creating two distinct tokenizations of the same text: the text as it appears on the papyrus ("original") and its editorially corrected ("regularised") version. Our tokenizer generates these versions utilizing the EpiDoc XML elements that denote editorial interventions (see Section 3.3 below for more details). Specifically, we classify all text as original except passages encoded as regularizations (`<reg>`), corrections (`<corr>`), additions (`<supplied>`) or deletions (`<surplus>`). For meaningful linguistic analysis, it is essential to pair each original word with its regularised counterpart. In the DDbDP documents, this is generally straightforward, as the `<reg>` and `<corr>` elements (alongside their corresponding original tags `<orig>` and `<sic>`) typically contain only a single word. However, more complex multi-word regularizations are also found in the source texts.[26] Our tokenizer addresses these with a custom string-matching algorithm that aligns the most similar words between the original and regularised elements.[27]

The tokenizer additionally collects various metadata for each word, including if it is a number (`<num>`), the section of the text (`<div>`), language (`<foreign>`), the scribe (`<handShift>`), the presence of expansions (`<ex>`) or additions (`<surplus>`), the line number (`<lb>`), and the sentence number. Identifying sentence boundaries is achieved simply by dividing the text at sentence-ending symbols, like periods and question marks (";" in Greek).[28] Finally, we store this metadata, along with the original and a (possible) regularised word form, in a MySQL database (see Section 3.4 below).

---

24 For instance, p.zauzich.39 (https://papyri.info/ddbdp/p.zauzich;;39, accessed 13.9.2023) contains extra spaces between letters (e.g. line 53: τ̣ο̣ υ̣ σ̣α̣τ̣αβο υς ο̣ι̣κ̣ ι̣[ας] should be τ̣ο̣υ̣ σ̣α̣τ̣αβους ο̣ι̣κ̣ι̣[ας]).

25 The complexity of the task is reflected in the fact that about 20% of the lines of code in our tokenizer deal with just these exceptions. The source code is available at https://github.com/erikhenriksson/papygreek-tokenizer.

26 For instance, in line 31 of pap.agon.3 (https://papyri.info/ddbdp/p.oxy;27;2476, accessed 15.8.2023) σεβαστάτῃ has been corrected to καὶ εὐσεβαστάτῃ.

27 Some texts contain more than one <reg> version; our tokenizer chooses the first one. However, we also mark the other alternatives and store them in the database for querying (the query implementation, however, is not yet ready as of this writing). Additionally, the texts contain variant readings encoded within `<lem>` and `<rdg>` elements. We choose to include the `<lem>` elements, since the `<rdg>` elements typically contain older and later corrected readings. However, as in the case of multiple `<reg>` versions, we also store the `<rdg>` tokens in our database.

28 These symbols, of course, are also editorial additions, representing modern interpretative decisions.

3.2 Morphosyntactic annotations

The PapyGreek database includes a selection of texts that we have morphosyntactically annotated using the Ancient Greek Dependency Treebank 2.0 annotation schema.[29] The annotation has been done using the Arethusa annotation environment,[30] which we have integrated into the PapyGreek platform. For the texts that are annotated, each token gets the following morphosyntactic data fields: 'lemma' (the dictionary form), 'postag' (the 9-character code encoding part-of-speech, person, number, etc.),[31] 'head' and 'relation'. The last two are part of the syntactic annotation schema, denoting the syntactic parent and the node's syntactic relation to it, respectively. We have annotated two versions of each sentence, the "original" and the "regularised" version (see Section 3.1 above). As of this writing, 650 texts have been annotated manually by experts in the Greek language. To ensure the quality of these annotations, PapyGreek has a review system where reviewers can accept or reject texts that annotators submit for review.

In addition to the manual annotations, PapyGreek also utilises automatic part-of-speech tagging and morphological analysis, based on the Ancient Greek BERT language model developed by Singh et al. (2021). The pre-trained model, available from Hugging Face[32], was fine-tuned for postagging using the AGDT, PROIEL and Gorman treebanks as training data,[33] and we further fine-tuned it using the PapyGreek Treebanks.[34] The postag prediction accuracy of the model for a holdout set was approximately 90%, which is acceptable considering the fragmentary nature of many of the included texts.[35] A limitation of the BERT model is that it can only predict POS tags but not lemmas. To include automatic lemmatization, we used the morphological analyser Morpheus[36] to map the predicted postag and its word form to its dictionary form. As with the manual annotations, we automatically annotated the original and regularised versions of the texts separately. Our search interface includes a confidence score (output by the BERT model) of the predicted POS tags and lemmas, allowing users to assess their quality.

29 https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/Greek_guidelines.md (accessed 25.8.2023).

30 https://github.com/alpheios-project/arethusa (accessed 25.8.2023).

31 For the full list of the POS tag codes used, see https://github.com/gcelano/LemmatizedAncientGreekXML (accessed 25.8.2023).

32 https://huggingface.co/pranaydeeps/Ancient-Greek-BERT (accessed 25.8.2023).

33 https://github.com/pranaydeeps/Ancient-Greek-BERT (accessed 25.8.2023).

34 Vierros – Henriksson 2021.

35 We plan to continue training the model as the training data (i.e. manually annotated treebanks) grows. The code for the automatic tagger can be found at https://github.com/erikhenriksson/papygreek-tagger.

36 We used the XML version of Morpheus, available from https://github.com/gcelano/LemmatizedAncientGreekXML/tree/master/Morpheus (accessed 25.8.2023).

3.3 Editorial interventions

Due to historical practices, editorial interventions in papyri are often labelled as "corrections". However, as Depauw – Stolk 2015 point out, the linguistically intriguing elements are typically the variant forms themselves, which can illuminate phonological and grammatical changes in everyday language usage.[37] Before the introduction of the Trismegistos Text Irregularities tool (see Section 2.1 above), there was no feasible method for systematically studying editorial alterations in the DDbDP as indicators of linguistic variation. For example, identifying every instance of ο being corrected to ω across more than 50,000 DDbDP source documents would have required navigating through each file's XML structure and running a string search within all elements potentially containing editorial changes – an exceedingly slow and inefficient operation. Trismegistos Text Irregularities pioneered fast access to this information by collecting these variations into a searchable database. PapyGreek Search adopts a similar approach, with some differences.

As outlined above (in Section 3.1), our tokenizer treats `<reg>` and `<corr>` elements as "regularised" text, and `<orig>` and `<sic>` as their "original" counterparts. Additionally, it utilizes editorial insertions (`<supplied>`) and deletions (`<surplus>`) to generate two versions of words that include these modifications. To make searchable the character-level differences between these versions, an appropriate algorithm was necessary. Finding differences between sequences is a well-known problem in computer science,[38] with several existing algorithms to choose from. We sought a method that would align well with our intuitions about the editors' intended corrections. Most character-based algorithms, by contrast, are designed to identify mathematically minimal edits between sequences, which are sometimes counterintuitive.[39] The algorithm we ultimately chose, Python's "difflib," utilizes gestalt pattern matching[40] to find the longest contiguous sequences that match in the compared strings, yielding results that seem to match human intuitions. Furthermore, difflib suggested more natural edits for word-initial differences than the other tools we evaluated. Take the word forms ενεγραφε and εγραφε. Both the "windiff" (Microsoft Windows) and "jsdiff"[41] interpret the difference as ε~~νε~~γραφε (remove the word-internal νε), while "difflib" gave the correct edit ~~εν~~εγραφε (remove the prefix εν).

37 See also Dickey 2011.

38 Gusfield 1997.

39 Myers 1986. For example, consider the two arbitrary sequences: (a) "to a blue table" and (b) "table". To a human, it is quite obvious that the difference is that (b) does not have "to a blue". However, most string-comparison libraries (e.g. the Unix "diff") suggest that to get from (a) to (b) one needs to remove all spaces, remove the character "o" from "to", remove "u" from "blue", respectively, and omit the final word "table".

40 Ratcliff – Metzener 1988.

41 By Kevin Decker, https://github.com/kpdecker/jsdiff.

After identifying the character-level edits (i.e. additions and removals) between the original and regularised word forms, as well as their left and right context, our system stores them in a separate MySQL database table. Importantly, we also include the parent word ID in each edit instance, which provides access to the data associated with the corresponding word (that is, lemma, postag, position in the sentence, etc.). This linking is crucial, since it enables queries combining linguistic variation with, for instance, part-of-speech (see Section 5 for example queries). The variant-collecting process is fully automated (in comparison, Depauw – Stolk 2015 describe a semi-manual process in creating the Textual Irregularities database).

3.4 Dependency relationships

One of the key goals – and technical challenges – of PapyGreek Search was enabling fast searches across morphosyntactic parsed texts In the AGDT schema we use (see Section 3.2 above), each word is associated with a single "head", its direct ancestor in the dependency tree. A sample XLM output of the Arethusa annotation environment is illustrated in Figure 1, showing how the head parameter references words in the same sentence.

```
<sentence id="1">
    <word id="1" form="ὃς" relation="SBJ" head="4"/>
    <word id="2" form="ἄν" relation="AuxY" head="4"/>
    <word id="3" form="τις" relation="ATR" head="1"/>
    <word id="4" form="εὕρῃ" relation="SBJ" head="9"/>
    <word id="5" form="τοῦτο" relation="ATR" head="7"/>
    <word id="6" form="τὸ" relation="ATR" head="7"/>
    <word id="7" form="ὄστρακον" relation="OBJ" head="4"/>
    <word id="8" form="," relation="AuxX" head="4"/>
    <word id="9" form="δώσει" relation="PRED" head="0"/>
    <word id="10" form="στατῆρα" relation="OBJ" head="9"/>
    <word id="11" form="." relation="AuxK" head="0" />
</sentence>
```

Fig. 1: An example output from Arethusa (simplified for demonstration by leaving out, among other things, the lemma and morphological information). Sentence 1 from o.claud.1.1, "regularised" version.

Similarly to the XML representation, our database stores the dependency relationships using the "head" column of the "word" table. Though simple and easy to manage, this format is ill-suited for querying purposes. It requires recursively rebuilding the tree structure for each queried sentence, a slow and computationally expensive process, especially when conducted across a large database of annotated sentences.

To address this, the PapyGreek Search MySQL database implements the so-called closure table design pattern.[42] This approach involves storing all possible paths within syntactic trees – that is, every child, grandchild, great-grandchild, and so on – in a separate database table. In other words, instead of storing just the *immediate ancestor* of each node (as in the traditional schema), we store *every descendant path* from each node down to the last non-branching child (see Fig. 2 for an illustration). This method simplifies queries and reduces query times by eliminating the need to reconstruct the hierarchy for each search separately. In addition, we link each stored path with the respective word records (using unique word IDs given to each word), which gives syntactic queries access to other word data such as its form, morphological annotation, and possible linguistic variations. This data linking forms the basis for the multi-domain search possibilities mentioned at the start (see also Section 3.6 below).

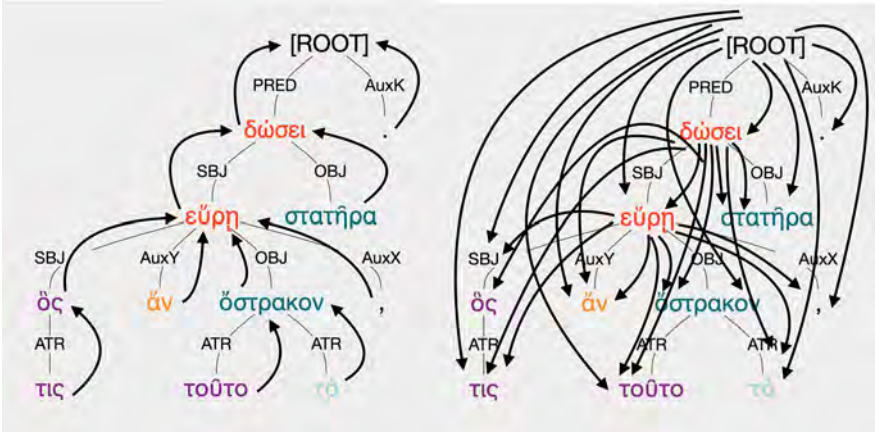


Fig. 2: Two syntactic dependency representation strategies: the conventional method, which records each node's immediate ancestor (left), vs. the closure table approach, capturing all nodes' descendant paths (right). The latter strategy takes up more space but makes for much more efficient querying across the hierarchy. Sentence 1 from o.claud.1.1, "regularised" version.

3.5 Metadata

In addition to storing data about words and sentences, the PapyGreek database also includes various document-level metadata. This is automatically collected from the DDbDP source files (including name, series name, main language, HGV numbers, TM numbers, last changed date) as well as from the linked HGV files (date not after, date not before, and place name). The date information is converted into integers (e.g. -300

42 Karwin 2010.

meaning 300 BC), which makes it easier to use them in queries. The place names, which in the source files are written with some variation (e.g. “oxyrhynchos”, “oxyrhynchus” and “oxyrynchos”), are normalized and mapped to *nomoi* (e.g. “Egypt: U19”) and other larger geographical names.[43] Users can choose to filter the search either using these converted names, or the original names as they stand in the HGV.

The PapyGreek system also includes manually entered metadata about the ancient people and genres associated with the texts. We split the documents by each <handShift> element and treat the generated splits as individual “acts of writing” (AOW).[44] Each AOW can be associated with one or more text types[45] as well as ancient persons playing different roles in the production of the document, including “author”, “writer”, “addressee”, and “external official”. Where the associated person has a Trismegistos Person ID,[46] we enter it in the metadata. Moreover, each person added to the PapyGreek database gets a unique identifier (“PapyGreek Person ID”). This is especially useful for distinguishing the same individual across multiple documents, even when the person has an unknown name and lacks a TM Person ID.

3.6 Relational database structure

The PapyGreek Search MySQL table structure comprises tables for preprocessed texts, separated into word and document tables, an edit table for character-level editorial changes within words, a closure table for syntactic structures, and metadata tables for persons and text types associated with different “acts of writing” (see Section 3.5 above). The key feature of this MySQL table structure are the links between data points (e.g., individual documents, words, syntactic relations, and variations), which enable combined searches using information from different tables. Figure 3 displays a simplified diagram of the relevant table relationships.

To illustrate these relationships, consider an example query. Suppose a researcher wants to identify the “original” word forms (see Section 3.3. above) lemmatised as εἰμί, specifically within syntactic constructions where the word has some subject as dependent. Additionally, she is only interested in those word forms where the editor has added a missing ε (e.g. forms like ἰμί corrected to εἰμί, ἰσίν corrected to εἰσίν, etc.), further restricting the search to documents dated to the 4th century AD.

43 The code for this conversion is available as part of the PapyGreek tokenizer (https://github.com/erikhenriksson/papygreek-tokenizer).

44 Vierros – Henriksson 2017.

45 We use a hierarchical text type list, with three levels (hypercategory, category, subcategory). For example, one of the hypercategories is called “private”. It includes the categories “letter”, “list”, “Memorandum (private)”, and “school text”. These categories further include subcategories; for instance, “letter” includes four possible subcategories, such as “private correspondence”.

46 Broux – Depauw 2015.

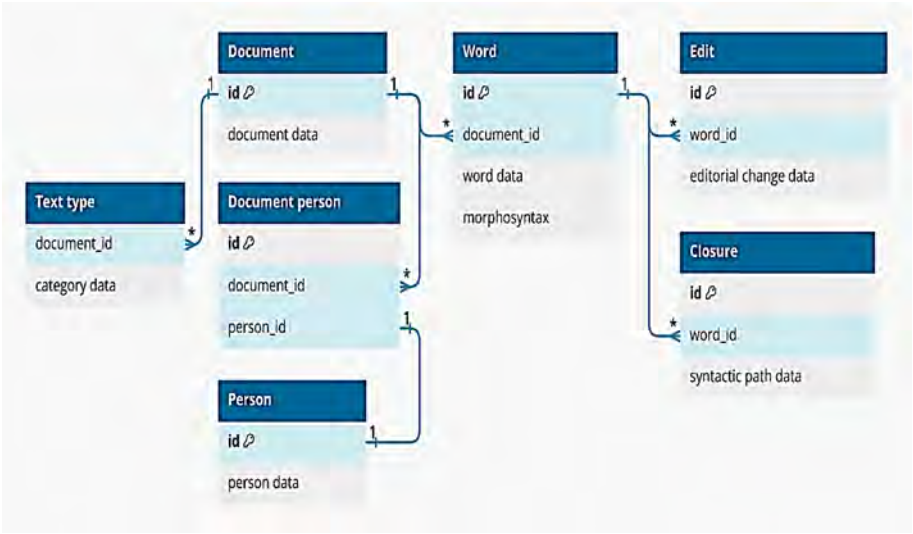


Fig. 3: A (simplified) diagram of the PapyGreek Search database, illustrating the relationships between tables that enable multi-domain queries.

Schematically, PapyGreek search processes the query as follows. First, the system identifies word records with the lemma εἰμί in the 'lemma_orig' field of the word table. Because each word is connected to its parent document, via this linking the retrieved words can be limited to documents that have the specified dates (300-399 AD) in their metadata fields. Next, using the closure table, which links syntactic dependencies to each word record, the search further narrows down to words having a subject dependent. Finally, the search selects only those words that meet the previous criteria *and* are present in the "edit" table with records containing an ε addition. This linkage of tables and gradual filtering of returned data is managed through unique identifiers associated with each database record.

This example query yields one result:[47] the word ἰσίν in the 6th sentence of p.mert.1.32, dated to the early 4th century. The sentence in question is ἰσίν τιναις γὰρ οἳ ἐπιθύμη[σαν] τοῦ τόπου, where ἰσίν indeed has a subject dependent (τιναις), as shown in Figure 4.

47 Retrieved 11.10.2023. Only one result was found because the treebank search targets the annotated sub-corpus in the PapyGreek database, which includes few documents from the 4th century. The query is built with two vertical search boxes (see Section 4.2 below for an explanation), with the parameters "lemma=εἰμί,form=+ε" in the upper box, and "relation=SBJ" in the lower box.

Fig. 4: Syntactic tree of ἰσίν τιναις γὰρ οἱ ἐπιθύμη[σαν] τοῦ τόπου from p.mert.1.32 (“original” version), showing the query result where ἰσίν has a subject dependent (τιναις).

4 User Interface

In this section, we walk through the main features of the PapyGreek search, beginning with an overview of the search interface. We then cover the available query parameters, followed by methods for identifying textual irregularities. Finally, we describe how to construct syntactic queries graphically.

4.1 Overview

The PapyGreek Search interface is divided into two main sections: 1) the search configuration section, where users set their search criteria, and 2) the results display section, where the outcomes of the search are presented. Starting with the search configuration section (Figure 5), users can first select to target either the “original” or “regularised” text versions (see Sections 3.1 and 3.2 for how these versions are created from the DDbDP source files).

Below this selection, a search box with a blue background allows users to target specific data fields of words stored in the PapyGreek database using various parameters, discussed below in Sections 4.2 and 4.3. Section 4.4 will further explain how to add more boxes to construct a syntactic tree search. The core functionality of the tool is centered around these search boxes: the parameters specified within them, and – in syntactic searches – the way they are arranged. Finally, the search area also includes several metadata fields, such as place, date, text type, and person, which can be used to further refine the search.

Fig. 5: A screenshot of the PapyGreek Search interface.

The results section (Figure 6) has three key elements. Firstly, it displays the results in a data table, where each retrieved word appears as a row accompanied by multiple data fields. These fields include links to the document on the PapyGreek platform, the sentence number and the word's position in it (showing its syntactic tree when clicked), the original and regularized forms of the word, morphosyntactic annotations, as well as date and place. The table includes automatic annotation data (see Section 3.2 above), along with the model's confidence score for them.

Secondly, the results section features a Timeline graph showing the number of results over 50-year intervals, with both absolute and relative frequencies displayed. Relative frequencies show the results as a percentage of the total word tokens in the DDbDP for each period. For the calculation, words from documents with uncertain dates spanning multiple 50-year intervals were distributed equally among these intervals.

Lastly, the interface allows exporting the results into a CSV file for further analysis outside the PapyGreek platform.

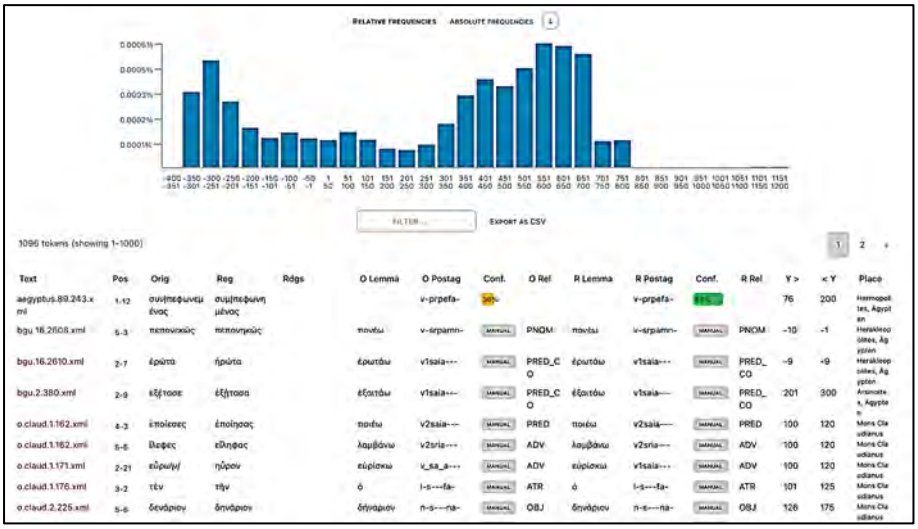


Fig. 6: Screenshot of the PapyGreek results section, displaying a data table of retrieved words and a Timeline graph showing results over 50-year intervals.

4.2 Query parameters

In the blue box located at the centre of the interface, query parameters are entered in the format: "parameter=value". The available parameters are the following: "form", "lemma", "lemma_plain", "postag", "relation", and "confidence".[48]

The `form` parameter is used for finding basic word forms and, using a special search syntax (discussed below in Section 4.3), also textual irregularities. This parameter ignores diacritics. For instance, to find word forms και in the uncorrected original text, one should select "Original" from the top version selection and enter "form=και" in the blue search box.

The parameters "lemma", "lemma_plain", "postag", and "relation" are utilized in morphosyntactic queries. The "lemma" and "lemma_plain" parameters allow for searching dictionary forms with and without diacritics, respectively. The "postag" parameter targets part-of-speech tags encoded as 9-character strings.[49] Automatic morphological annotations (see Section 3.2 above) can be optionally filtered out by using the "confidence" parameter within the range 0-1. For example, to require a score of 0.9 or above (indicating a high confidence), one would use "confidence=>0.9"; setting the value

48 In addition, there are two parameters, "order" and "depth", which are used in treebank queries and will be discussed below in Section 4.3.

49 For the list of available part-of-speech tag characters and their meaning, see https://github.com/gcelano/LemmatizedAncientGreekXML (accessed 12.9.2023).

to 1 is a shorthand for choosing only manual annotations. Lastly, the "relation" parameter targets syntactic relations annotated in the AGDT 2.0 scheme (see Section 3.2 above). To identify predicates, for instance, one would enter "relation=PRED".

The wildcard symbol % and underscore _ have a special meaning in the parameters listed above. % can replace any string of characters, while _ is used to skip a single character. For example, entering "form=κ_μη%" would return forms like κώμης and καμήλια. These symbols can be particularly useful with part-of-speech tags, such as in the query "postag=v%" to locate all verbs. As an alternative method for constructing complex queries, each parameter can optionally operate with regular expressions.[50] It can be used by adding the prefix "regex:" to the parameter name.

The parameters can be combined with a comma; for example, one can search for adverbial adjectives by typing "relation=ADV,postag=a%" in the search box.

4.3 Variation search

The linguistic variant search feature of PapyGreek Search utilizes the "form" parameter with four special symbols: + (plus) for editorial additions, - (minus) for deletions, > for left-hand context and < for right-hand context.[51] For instance, searching with "form=-ι" identifies all occurrences where an editor has deleted ι, whereas "form=+ι" finds cases of ι being added. For editorial replacements, a combination of - and + symbols is used. For example, to find instances where ε has been replaced with η, one would enter "form=-ε+η". The symbols > (before) and < (after) symbols are used to narrow down the search by word context. For example, "form=λ>-ε+αι" targets corrections from ε to αι following λ (in the same word).

For more precise criteria, textual variant queries can be further refined using regular expressions via the "regex:form=" parameter. When using the regex mode in this context, each of the included subqueries (for example, the parts after + or before <) must be determined with their own regex search pattern.[52] As an illustration, suppose one wishes to find all cases ω or ωι corrected to ου, but only when this correction appears word-finally. Using regex, this could be constructed as the following query: "regex:form=-^(ω|ωι)$+^ου$<^$".

The query consists of three parts: the string removed by the editor (-^(ω|ωι)$), the string that was added in its place (+^ου$) and the empty right-hand context (<^$). The |

50 E.g. Goyvaerts – Levithan 2012.

51 When conducting variation searches, the choice between "Original" and "Regularised" at the top of the interface is ignored.

52 Note that the symbols + and - are already reserved in the variation syntax and so cannot be used as quantifiers in the regex pattern. As a workaround, one can use the symbols ＋ (Full-width Plus, U+FF0B) and － (Full-width Hyphen-minus, U+FF0D). The PapyGreek Search system will transform these back into their normal regex counterparts + and -.

symbol denotes alternatives, and so ω|ωι means ω or ωι. The ^ and $ indicate beginnings and ends of strings, respectively, and are used here to restrict the search to exactly the given substrings. For example, by typing +^ου instead of +^ου$, the query would match all editorial additions *starting* with ου. Finally, the <^$ part specifies the empty word-final context.

4.4 Treebank search

PapyGreek Search has a graphical interface for building parametrized subtrees to explore syntactic dependencies. Users can add more search boxes to their query by clicking the "+" button at the bottom of an existing search box, which adds a new box underneath. This allows users to include as many boxes as needed to construct the desired tree structure.

In each box, any of the previously mentioned parameters can be used, including the special variation symbols detailed in Section 4.3 above. For example, to find (sub)trees where an object and some adverbial depend on a predicate, one would enter "relation=PRED" in the first box and add another box below it with "relation=OBJ". Then, one would add another box below the first box with the input "relation=ADV". This query is illustrated in Figure 7. Search boxes can also be moved horizontally using the left and right arrows found within each box.

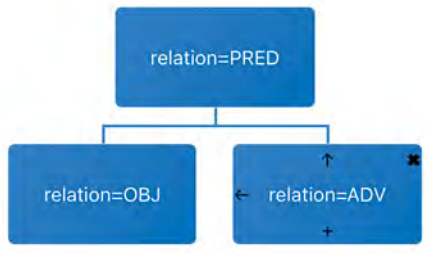


Fig. 7: Example query to find (sub)trees with an object and adverbial depending on a predicate.

The "depth" parameter enables specifying the tree depth in relation to the parent node, defaulting to 1 to indicate an immediate subordinate position. This parameter can be set to a numerical value, like 2, to skip a level, or an asterisk (*), to ignore depth and match all nodes under the specified parent.

Additionally, treebank queries can be refined by word order using the "order" parameter, which accepts any integer (e.g., 1, 2, 3, 4). This number indicates the word's relative position in the sentence, with lower numbers representing earlier positions. For instance, "relation=PRED,order=2" in one node, and "relation=OBJ,order=1" in another would match sentences where the predicate follows the object, as 2 > 1.

5 Example queries

This section presents brief examples of queries using PapyGreek Search, intended to illustrate the kind of linguistic research the tool can facilitate. First, we explore case variation in definite articles. Then, we address the representation of female protagonists in the papyri through the use of personal pronouns.

5.1 Case confusion or spelling error?

In another chapter of this book (Henriksson – Dahlgren – Vierros), we explore a case study on vowel variation in Greek related to /o, u/ allophonic variation in Egyptian. We observe that variations between ο /o/, ω /o(:)/, and ου /u/ often occur at the end of words, and note that the traditional explanation for case endings that contain these vowels involves the morphological merger of the genitive and dative cases. Here, we may take a different angle to this issue by examining noun phrases containing a definite article. If all words in the noun phrase follow the same case inflection, it lends more support to the case merger explanation, even though allophonic variation may have facilitated the merger.[53] We can also see if similar case variation occurs in feminine or plural noun phrases that contain different vowels. The singular masculine genitive and dative cases of the definite article, τοῦ /tu/ and τῷ/τῶι[54] /to(:)/, are relevant here, with identical forms in the neuter gender. In the feminine singular, the article is τῆς /te(:)s/ or /tis/ (gen) and τῇ /te(:)/ or /ti/ (dat). The accusative singular form adds another flavour to this soup.[55]

We can formulate a query for the definite article either using the lemma (ὁ for all genders), or the postag for the article, singular and masculine (l_s__m%), and combine it with the spelling variation as ου corrected by the editor to ω (-ου+ω).[56] This yields masculine articles where the original text is τοῦ and the regularized form is τῷ, resulting in 58 hits – a relatively low number. The same query for the neuter[57] adds another 22 hits. It is more convenient to combine the masculine and neuter queries by using the regular expression option, allowing us to include the optional iota adscript written in the original papyrus, without needing to specify gender at all in the postag.[58] When we

53 See, e.g., Stolk 2015 concerning the similarities of the semantic roles of the genitive and the dative affecting the case merger.

54 The *iota* in the dative ending was silent; in the papyri it was either not written at all (τω) or it was written as adscript, after the *omega* (τωι); the marking the *iota* below the *omega* as a subscript (τῳ) is merely an editorial habit that has no significance in respect to what was written on the papyrus. (Clarysse 1976; Horrocks 2010, 116; Vierros 2012, 121–36).

55 Horrocks 2010, 116.

56 https://papygreek.com/search/192 (accessed 28.2.2024).

57 Replacing 'm' with 'n' in the postag string.

58 Search input: "postag=l_s%,regex:form=-(ω|ωι)+ου".

change the spelling the other way around (+ου-(ω|ωι)) we get 120 hits, but not all of them are singular genitives in the regularised text; the results include also instances such as τοῦ corrected into τῶν, the plural genitive form. This can be avoided by adding a word-final position restriction (-(ω|ωι)<^$), but it is good to keep in mind also these forms concerning the variation between the vowels ου and ω. After the restriction for word-final position, we have 84 instances of τοῦ written instead of τῶ/τῶι.

How do we get to querying the noun phrases, then? As explained in Section 4.4 above, this is simple by adding more search boxes through the search interface. In one box, representing the head, we search for a word in a specific case, and in another box, representing the dependent, we specify its article, including any potential spelling variants. Since the treebanked corpus is small, we don't yet see many results for these queries. For example, a query for a head in the genitive case with a definite article as its modifier that has been written with a dative ending (Figure 8) yields only one result, upz.1.79, sentence 7 (l. 5): τω Ἄμμωνος, where the definite article looks like a dative, whereas the name of the god Ammon is in the genitive case. The genitive case would be correct here, but the two preceding words are correctly in the dative singular (the whole phrase being ἐν τῷ οἴκῳ τοῦ Ἄμμωνος "in the house of Ammon"). The writer may have written the ω in the definite article as a continuation of the cases from the previous words, or they may have aimed to write the genitive case but 'failed' because of its phonetic similarity with the dative case.

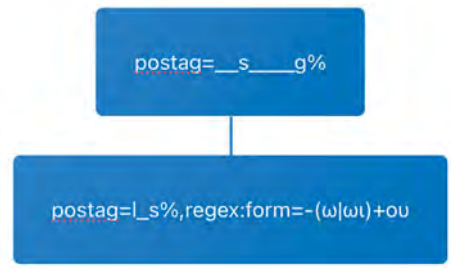


Fig. 8: A query with a head word in genitive singular and its dependent, an article in singular presenting spelling ω/ωι in the original papyrus that has been regularized into ου.

While waiting for the treebanked corpus to grow, we might continue this case study by reviewing the results for queries on article spelling variations and analyzing their linguistic contexts – does the article and the head have the same case, or is the nonstandard case of the article in disagreement with the head, and so forth. For instance, one of the first results from the article query above is bgu.3.998, which presents a similar case as the tree query in Fig. 8, τω τούτου υἱοῦ Ἁρπαήσιος (instead of τοῦ). The correct genitive with ου would definitely be expected also by analogy, since two out of three of the following genitives in the same noun phrase have those letters.

Repeating the query process for feminine articles and their head words would guide us further in in determining whether phonetics primarily drives the observed variation, or if the cases are indeed understood and used as semantically merged.

5.2 Gender by way of pronouns

Sociolinguistically and historically, there is interest in understanding the ways in which female protagonists are acting in the papyrological corpus. In addition to identifying female actors in papyri from prosopographical information provided by Trismegistos People or from the author and addressee metadata in the PapyGreek corpus, for example, we can also examine the use of personal pronouns using PapyGreek Search to understand the roles these pronouns play in the linguistic context.

For this preliminary study, we focus on the first-person pronoun. A basic lemma query ("lemma=ἐγώ") would retrieve all cases and genders of the 1st person pronoun. The same could be achieved with a postag query, which can also be further refined to search for specific genders and cases ("postag=p1s___f%" for pronoun, 1st person, singular, feminine). The search for feminine pronouns[59] brings us 243 hits. However, a caveat here is that this result does not present the whole corpus of papyri, but only the manually annotated part, where annotators have been able to mark as feminine those pronouns that actually do refer to women. In the Greek language, first and second person pronouns do not have gender-specific forms; this distinction only exists in third-person pronouns. The masculine bias in the training corpus is evident, as the automatic tagger has not assigned feminine gender to *any* first-person pronoun.[60] To illustrate this discrepancy, a query for masculine singular first-person pronouns gives 5296 hits, covering both manually and automatically tagged pronouns. So, we must continue our case study with the 243 instances where the feminine gender has been manually marked. The role these persons play in each text can already be considered important, as they represent the 'me' in the text. However, we can further look at the syntactic role of the pronoun either by using the relation tag (see Figure 9) or simply by using the grammatical case (nominative often expressing the subject, accusative usually being the case of the direct object and dative as the case of the indirect object or recipient). The relation tag, however, can be seen as more precise in identifying the actual roles of the pronouns. For instance, the subject might appear in cases other than the nominative, such as the accusative in *accusativus cum infinitivo* structures or the genitive in genitive absolute con-

59 https://papygreek.com/search/190 (accessed 28.2.2024).

60 One can question the linguistic correctness of marking the gender in the annotated corpus for pronouns that do not bear the marker for gender. However, in the annotation process we also select to mark other features requiring interpretation, like the grammatical case of neuter accusatives and nominatives, which look alike. In addition, this marking can help in research as we try to show here, as long as we are aware of the basis and procedure of the annotation.

structions, and cases can also be determined by other factors, such as prepositions. The distribution of the syntactic relations of the feminine first person pronouns is shown in Figure 10. In over half of the occurrences the 'me' in the text is the object, and more specifically the indirect object, since from the 124 instances of OBJ, 96 are in the dative case. The role as an attribute is given in roughly a third of the cases, which can indicate, for example, a possessive role marked by the genitive case. The low frequency of the subject role is not as alarming as it may sound, given that Greek embeds subject information within the verb's inflection and does not always use personal pronouns to express the subject. To query female subjects of verbs, we would also need the ability to combine female personal names from previous sentences, where their role as the subject is clear, along with possibly other methods.

```
lemma=ἐγώ,postag=p1s___f%,relation=SBJ%
```

Fig. 9: Query for the first person singular pronoun with the relation subject (including coordinated subjects).

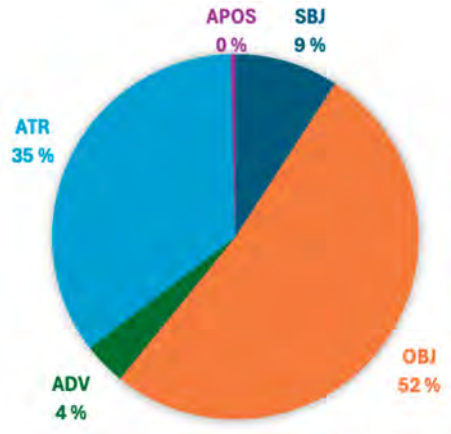


Fig. 10: Syntactic relations of the feminine first person pronouns.

6 Conclusion

We have introduced PapyGreek Search, an online open-access query tool for documentary Greek papyri. PapyGreek Search stands out for its integrated search interface, en-

abling simultaneous queries across linguistic domains, such as morphology, syntax, and, indirectly through editorial interventions, phonological and morphosyntactic variations. The treebank search feature of PapyGreek Search includes a graphical interface for query construction, making it accessible to non-technical users. In this chapter, we outlined the motivation for creating the tool, described its implementation in detail, and provided a walkthrough of its main features. Additionally, we demonstrated how the tool can be used for linguistic research with two example queries, addressing potential case confusions and the use of genders in personal pronouns.

Bibliography

Baumann, R. (2013), *The Son of Suda On-Line*, in *The Digital Classicist*, ed. by S. Dunn – S. Mahony, https://ryanfb.xyz/papers-BICS/sosol-bics-draft.pdf.

Broux, Y. – Depauw, M. (2015), *Developing Onomastic Gazetteers and Prosopographies for the Ancient World Through Named Entity Recognition and Graph Visualization: Some Examples from Trismegistos People*, Social Informatics 8852, 304–313.

Clarysse, W. (1976), *Notes On The Use Of The Iota Adscript In The Third Century B.C.*, CdE 51, 150–66.

Depauw, M. – Keersmaekers, A. (2017), *Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri*, Classics@ 20, https://classics-at.chs.harvard.edu/bringing-together-linguistics-and-social-history-in-automated-text-analysis-of-greek-papyri.

Depauw, M. – Stolk, J. (2015), *Linguistic Variation in Greek Papyri: Towards a New Tool for Quantitative Study*, GRBS 55, 196–220.

Dickey, E. (2011), *The Greek and Latin Languages in the Papyri*, in *The Oxford Handbook of Papyrology*, ed. by R. Bagnall, Oxford, 149–69.

Evans, T. V. – Obbink, D. D. (2010), *The Language of the Papyri*, Oxford.

Gignac, F. T. (1976), *A Grammar of the Greek Papyri of the Roman and Byzantine Periods, Volume 1: Phonology*, Milan.

Goyvaerts, J. – Levithan, S. (2012), *Regular Expressions Cookbook*, Sebastopol (CA).

Gusfield, D. (1997), *Algorithms on Strings, Trees, and Sequences*, Cambridge.

Horrocks, G.C. (2010), *Greek: A History of the Language and Its Speakers*, 2nd ed., Oxford.

Karwin, B. (2010), *SQL Antipatterns: Avoiding the Pitfalls of Database Programming*, Pragmatic Bookshelf.

Keersmaekers, A. – Van Hal, T. (2023), *Creating a Large-Scale Diachronic Corpus Resource: Automated Parsing in the Greek Papyri (and Beyond)*, Natural Language Engineering 2023, 1–30.

Keersmaekers, A. – Mercelis, W. – Swaelens, C. – Van Hal, T. (2019), *Creating, Enriching and Valorizing Treebanks of Ancient Greek*, "TLT, SyntaxFest" 2019, 109–17.

Krause, T. – Zeldes, A. (2016), *ANNIS3: A New Architecture for Generic Corpus Query and Visualization*, Digital Scholarship in the Humanities 31, 118–39.

Martens, S. (2013), *TüNDRA: A Web Application for Treebank Search and Visualization*, in *Proceedings of The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, Sofia, 133–44.

Mayser, E. – Schmoll, H. (1970), *Grammatik Der Griechischen Papyri Aus Der Ptolemäerzeit. Laut- Und Wortlehre: Einleitung Und Lautlehre*, Berlin.

Van Minnen, P. (1996), *The Duke Data Bank of Documentary Papyri*, https://library.duke.edu/papyrus/texts/DDBDP-old.html.

Myers, E. W. (1986), *An O(ND) Difference Algorithm and Its Variations*, Algorithmica 1, 251–66

Pajas, P. – Štěpánek, J. (2009), *System for Querying Syntactically Annotated Corpora*, in *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, Singapore, 33–6.

Ratcliff, J. W. – Metzener, D. (1988), *Pattern Matching: The Gestalt Approach*, Dr. Dobb's Journal, July 1988, 46.

Rosén. V. *et al*. (2017), *Exploring Treebanks with INESS Search*, in *Proceedings of the 21st Nordic Conference on Computational Linguistics*, Gothenburg, 326–9.

Singh, P. – Rutten, G. – Lefever, E. (2021), *A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek*, in *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 128–37.

Sosin, J. (2010), *Digital Papyrologyi*, https://blog.stoa.org/archives/1263.

Stolk, J. V. (2015), *Dative by Genitive Replacement in the Greek Language of the Papyri: A Diachronic Account of Case Semantics*, Journal of Greek Linguistics 15, 91–121. [https://doi.org/10.1163/15699846-01501001]

Vierros, M. (2012), *Bilingual Notaries in Hellenistic Egypt: A Study of Greek as a Second Language*, Brussel.

Vierros, M. (2018), *Linguistic Annotation of the Digital Papyrological Corpus: Sematia*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 105–18.

Vierros, M. – Henriksson, E. (2017), *Preprocessing Greek Papyri for Linguistic Annotation*, in *Journal of Data Mining and Digital Humanities. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*, ed. by M. Büchler – L. Mellerin, http://jdmdh.episciences.org/paper/view/id/1385.

Vierros, M. – Henriksson, E. (2021), *PapyGreek Treebanks: A Dataset of Linguistically Annotated Greek Documentary Papyri*, Journal of Open Humanities Data 7, https://doi.org/10.5334/johd.55.

Vierros, M. – Yordanova, P. (2022), *Querying Syntactic Constructions in Ancient Greek Parsed Corpora: A Case Study on the Genitive Absolute in Literature and Documentary Papyri*, Classics@ 20, https://classics-at.chs.harvard.edu/querying-syntactic-constructions-in-ancient-greek-parsed-corpora-a-case-study-on-the-genitive-absolute-in-literature-and-documentary-papyri.

Erik Henriksson — Sonja Dahlgren — Marja Vierros

Phonological Variation in Greek Papyri

Two Case Studies Using PapyGreek Search

1 Introduction

Documentary papyri[1] have long been recognized as a remarkably direct source of insight into Postclassical Greek. From a linguistic perspective, these texts are invaluable for the various types of writing errors they contain, which can offer significant clues about the evolution of Greek during the Greco-Roman period.[2] Earlier, papyrologists had to rely on traditional reference texts such as those by Mayser and Schmoll and Gignac[3] to interpret these irregularities. However, digital advancements have reshaped the field,[4] and now that all published texts are digitized and accessible online (https://papyri.info), linguistic variants are accessible in a machine-readable format, enabling new methods of analysis.

In this study, we delve into the phonological aspects of linguistic variation in Greek documentary papyri using PapyGreek Search, a query tool developed as part of the PapyGreek project (https://papygreek.com/search).[5] Our focus is on the phonological variation that seems to be related to the prolonged contact between Greek and Egyptian-Coptic – a language contact situation that lasted for over a millennium after Alexander the Great's conquest of Egypt. Given that most papyrus findings originate from Egypt, the interaction between Egyptian (in its Demotic or Coptic stages) and Greek used in Egypt is relatively well-documented. Frequently, Greek scribes were Egyptians with varying levels of proficiency in Greek as a second language, leading to distinct errors in writing.[6] As Gignac first noted in 1976 and reaffirmed in 1991, many unconventional spellings in documentary Greek could be attributed to phonological transfer from Egyptian-Coptic due to bilingualism.[7] Utilizing our novel tool, we explore this interaction further, building on previous studies such as Dahlgren 2017 and 2020.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 758481).

1 Following common practice, the term 'papyri' in this study will also include ostraca and wooden tablets.
2 E.g. Dickey 2011.
3 Mayser – Schmoll 1970; Gignac 1976.
4 Reggiani 2017.
5 Henriksson – Vierros, this volume.
6 E.g. Vierros 2012.
7 Gignac 1976, 57–60; Gignac 1991.

https://doi.org/10.1515/9783111070162-012

The chapter proceeds as follows. Section 2 reviews the research tools available for linguists interested in studying phonological variation in Greek documentary papyri (2.1). We also discuss the main types of phonological variation found in the texts, with a focus on the particular contact features typical to Egyptian Greek, such as underdifferentiation of phonemes and coarticulation (2.2 and 2.3). Section 3 details our methodology, including data collection and the implementation of an algorithm for finding the character-level differences between linguistic variants. Our case studies are presented and evaluated in Section 4. In Section 5, we discuss our findings, offering preliminary conclusions and directions for future research.

2 Background

In this section, we first review the research tools available for linguists interested in studying phonological variation in Greek documentary papyri. We then present the primary phonological variations found in these materials, focusing on those possibly influenced by language contact. Finally, we discuss coarticulation, a recurring characteristic of the Egyptian variety of Greek.

2.1 Previous work and PapyGreek Search

Past research into the phonology of Greek documentary papyri is primarily found within two significant works: the phonology volume of Edwin Mayser's *Grammatik der Griechischen Papyri aus der Ptolemäerzeit* (1926–1938), updated by H. Schmoll,[8] and Francis T. Gignac's *Grammar of the Greek Papyri from the Roman and Byzantine Periods* volume 1.[9] These grammars offer extensive, manually compiled lists of orthographic variations in the texts, a practice also followed by Teodorsson 1977. While these resources remain valuable to papyrologists, they have notable limitations. First, their approach is mainly qualitative and selective, offering only broad statistical insights.[10] Furthermore, e.g. Gignac 1976 is overly cautious in labelling nonstandard features to result from language contact; Gignac 1991 states this more clearly but is lesser known as a source to explain the variation. Also, with the significant increase in published papyri since the 1970s, some of the earlier data and interpretations have inevitably become obsolete.

8 Mayser – Schmoll 1970.

9 Gignac 1976.

10 Gignac 1976, for instance, offers statistical information rather vaguely using vocabulary such as "sometimes", "occasionally", "frequently", "very frequently" throughout his work.

Turning to more recent work, the Trismegistos Text Irregularities database (TMTI)[11] emerged as the first online platform for examining linguistic variation within Greek documentary papyri. The search functionalities of TMTI mirror those of the PapyGreek Search's linguistic variation search features. Both tools, for instance, allow users to search for character-level editorial modifications, including deletions, additions, and replacements.[12] Furthermore, both allow users to tailor their searches based on context, such as the corrections of ε to η either before or after κ. However, the tools also differ in some important respects. Firstly, the TMTI platform, as detailed in its "Methodology" section online,[13] relies on data collected in 2014 and 2016, while PapyGreek Search takes advantage of the most recent editions, updated weekly from the papyri.info data repository (https://github.com/papyri). Moreover, PapyGreek Search provides a higher degree of flexibility than TMTI, which currently restricts users to a context of only one letter or a diphthong before and after the correction. In PapyGreek, the query context can be placed *anywhere* within the word, rather than strictly before or after the changed letters. This feature, along with several other options for refining the search, is made possible by the integration of Regular Expressions[14] into PapyGreek Search's parameters. Finally, PapyGreek Search features a timeline chart of the search results, facilitating the exploration of diachronic trends of the phenomenon under investigation (see Section 4 for examples). For a closer look at PapyGreek Search and its functionalities, see Section 3 below and Henriksson – Vierros, this volume.

2.2 Types of variation

Much of the variation found in Egyptian Greek is similar to the variation common in Koine Greek everywhere, related to internal Greek phonological processes. These processes led, for instance, to many vowel qualities fronting and/or raising to [i], which is the result in Modern Greek. However, Egyptian Greek had some peculiarities that were not found elsewhere. These include the underdifferentiation of the voiceless and voiced stops (/p, t, k/, /b, d, g/), which frequently resulted in having a voiceless variant where a voiced one should be, and vice versa, probably often due to hypercorrection. Such phenomena stem from the absence of the voiced stops /g, d/ in Coptic (Loprieno 1995; Layton 2000). Nevertheless, Egyptian scribes knew that they existed in the Greek they used for

11 See Depauw – Stolk 2015.

12 As Depauw – Stolk 2015, 212–6 point out, editorial practices have not been consistent in the long history of editing and publishing Greek papyri. Consequently, despite more recent efforts of standardization, some textual irregularities have not been encoded as such in the digitized texts, and are not detected by either TMTI or PapyGreek Search. An algorithm that would find these missing variants would be desirable.

13 https://www.trismegistos.org/textirregularities/methodology.php.

14 Aho 1991.

work, and occasionally added them where they did not belong because it seemed 'Greek'. Consider, for instance, the variation between /g/ and /k/ in Table 1. As can be seen from the examples, there are misspellings regarding this particular underdifferentiation of phonemes from the Ptolemaic period to quite late into the Roman period. In all, there are 566 instances in the PapyGreek database of γ /g/ being replaced with κ /k/, and 1.489 vice versa.[15] However, the many instances of standard /k/ having been replaced with /g/ are partly explained by the both Greek and Egyptian-Coptic ten-dencies of assimilating voiceless consonants to voiced ones, as the next word after εγ was often δίκης or δεξιῶν, and it is likely that the scribes learned the form as a semi-standard (Dahlgren 2017, p. 155). In any case, confusion between voiced and voiceless stops was not a regular feature of Greek language-internally, so the phenomenon in Egyptian Greek can be connected to the language contact situation.

Table 1: Examples of variation between /g/ and /k/.

Nonstandard	Standard	Document	Date	Provenance
κεωρκων	γεωργῶν	o.narm.5	150–225 AD	Arsinoites
κρυτωπωλων	γρυτοπωλῶν	bgu.1.9	276 AD	Arsinoites
θυκατηρ	θυγάτηρ	o.heid.332	101-200 AD	Thebes
μετηνεκκα	μετήνεγκα	upz.1.14	157 BC	Memphis
εκραφη	ἐγράφη	stud.pal.3.303	617 AD	Arsinoites
εγ	ἐκ	p.amh.2.46	113 BC	Pathyris

Greek used in Egypt also had other types of variation particular to the region. Some variations involved nonstandard markings of consonants, such as lambdacism, i.e., variation between the two liquid consonants /r/ and /l/, due again to (Egyptian-)Coptic not having /r/ at least in all dialects.[16] The quality of /r/ in Egyptian-Coptic is a topic we will explore in this chapter (Section 4.1). Frequently, however, the variation is concerned with vowel orthography, which seems to have had a mind of its own in Egyptian Greek and has been studied far and wide by both Greek scholars and Coptologists alike.[17] The vowel variation concerned practically all interchanges between e.g. /ai/ and /e, a/, /i/ and /e:, ei, y, oi/, /u/ and /y/, as well as between /o/ and /u/. This last case was again linked to the language contact situation but was related to the different stress patterns of Greek and Egyptian-Coptic, rather than phoneme inventories. We study this topic in Section 4.2.

15 The search parameters used to find variations between /k/ and /g/ are "form=-κ+γ" and "form=-γ+κ". The searches were conducted on 1st July 2023. (For an explanation of PapyGreek Search's query syntax, see Section 3.4 below.)

16 Peust 1999, 127–32.

17 Girgis 1966; Gignac 1976; Teodorsson 1977; Consani 1993; Torallas Tovar 2010; Horrocks 2010; Dahlgren 2017.

Much of this variation, naturally, is a product of the internal development of Greek. The phenomenon known as iotacism, or itacism, which resulted in the Greek front and close vowels merging into /i/ over time, certainly caused variation. This is evident not only in documents penned by native Greek (L1) writers but also those written by second language (L2) users. This would, of course, be related to such matters as bilingual language users or those with at least some competency in Greek becoming accustomed to hearing especially high-frequency words with the phoneme /i/, even when they had been previously pronounced with /e:/ or /ei/, as in the word ἐκεῖνος. This could result in nonstandard orthographic forms, such as those shown in Table 2, if the writer did not remember the orthographic standard.

Table 2: Example of variation related to iotacism.

Nonstandard	Standard	Document	Date	Provenance
εκινος	ἐκεῖνος	p.col.8.242	401–500 AD	Arsinoites

The papyrological documents, therefore, offer a rich body of evidence that can, using tools such as PapyGreek Search[18] or Trismegistos Text Irregularities,[19] be studied with the intent of pinpointing certain phases of the diachronic development of Greek. For instance, we can seek to establish the time when the variation between /i/ and /ei/ began to surge, suggesting a merger of these two phonemes, or when exactly the former Greek stops /d/ and /b/ started to develop into the voiced fricatives /ð/ and /β/ that we observe in Modern Greek.[20] Additionally, the search tools can be used to study the other language involved in the scenario, giving clues about its properties through the misspellings of the L2 Greek users. And, as much as there was variation seemingly related to iotacism, even that was often tied to the strongest element of Coptic phonology, consonant-to-vowel coarticulation.

2.3 Coarticulation

Coarticulation, a phenomenon studied in articulatory and acoustic phonetics, refers to the process of sounds adapting to the quality of the adjacent ones in continuous speech.[21] It can either be anticipatory or carryover, or both; this means that sounds can either affect the following sounds or that there can be phonetic residues of the previous sounds still remaining on the sounds that follow. For example, a vowel can be changed

18 Henriksson – Vierros, this volume.
19 Depauw – Stolk 2015.
20 Horrocks 2010, 112.
21 E.g. Hardcastle – Hewlett 2000.

in quality by the uvular consonant /q/, which is produced further back within the vocal tract, so that /i/ results in a retracted production similar to [e].

Coarticulation is part of all speech; it is impossible to speak without sounds overlapping one another to some extent. But coarticulation can be very language-specific, serving the distinctive needs of languages regarding their most important phonological contrasts.[22] As we know, in language contact, the L1 features of a language are often transferred onto the L2 used.[23] This includes L1 coarticulatory patterns. Frequently, when studying features of a contact language variety, the research ends up revealing as much or even more about the L1 features causing the variation than about any new patterns within the contact language itself. Dahlgren's 2020 study of Coptic vowel reduction, based on L2 Greek misspellings, serves as an example.[24]

In this study, we focus on coarticulation because it is a distinctly Egyptian-Coptic feature.[25] Coarticulation is not part of Greek diachronic development to the same extent in general, and more specifically, it is not similar in Greek as in (Egyptian-)Coptic, i.e. vowels do not adapt according to the consonantal context in the same phonetically systematic way as they do in (Egyptian-)Coptic. Egyptian-Coptic as a language belongs to the Afroasiatic language family, which has many other examples of the same type of consonant-to-vowel coarticulation in e.g. Arabic.[26] It is therefore a specific element that differentiates Egyptian Greek as a contact variety from the Greek-internal developments that formed the basis of the Koine Greek form used in all Greek-speaking areas from the Roman Period onward.[27]

3 Methodology

This section outlines the methodology used in our case studies examining phonological variation in documentary Greek. We begin by offering an overview of the data collection and preprocessing procedures utilized in PapyGreek Search. This is followed by a description of our method for identifying linguistic variants and the associated algorithms used for discerning character-level differences. Finally, we give a brief overview of our methods for storing and querying these variants.

22 Manuel 1999.
23 Weinreich 1968; Thomason – Kaufman 1988; Matras 2009 etc.
24 Dahlgren 2020.
25 Dahlgren 2017.
26 Bellem 2007; Ryding 2005.
27 Dahlgren 2022.

3.1 Data source and preparation

We use the texts from the Duke Databank of Documentary Papyri (DDbDP), a comprehensive digital repository containing more than 50,000 documentary papyri, accessible at https://github.com/papyri/idp.data. The encoding of these texts is based on the EpiDoc XML schema,[28] a system that utilizes various XML tags to manage the transcriptions and editorial handling of the texts. Our text preprocessing methodology is delineated by Henriksson – Vierros in this volume;[29] here it suffices to note that our linguistic variation database hinges on the modern editorial corrections embedded in the texts.[30] As an example, an irregular form like διδι ("give") might be regularized by an editor to διδε. These variant forms are contained by `<choice>` tags in the XML schema, with `<orig>` and `<reg>` elements nested within the parent tag representing the original and standardized words, respectively. What we are interested in are the character-level differences between the original and standardized forms (e.g., the "ι" in διδι corrected to "ε" in διδε), including the surrounding context of those differences. Therefore, we needed an algorithm capable of detecting these differences between the variant forms.

3.2 Identifying character-level variants

The search for a suitable algorithm for the task at hand required a comparison of various options. Our primary requirement was to find an algorithm that could find character-level differences which correspond to the editor's intended corrections, which is sometimes different from detecting *minimal* string differences from a computational perspective.[31] We also considered how different tools respond to ambiguous differences. For example, given the variants εγραφε and ενεγραφε, it appears evident to a human that the difference lies in the prefix εν-, which is absent from εγραφε. Yet, many tools would interpret this as the addition of an infix νε after the initial ε, which is clearly not the case here. After examining numerous algorithms, we found that Python's "difflib" library, utilizing gestalt pattern matching,[32] offered the most intuitive results. Therefore, we chose to use this library to find the character-level edits. For details of the process, see Section 3.3 in Henriksson – Vierros, this volume.

Our system treats the differences between the original and regularized forms as a series of *edit instructions*, essentially a guide on how to transform the original form into its corrected version. Four types of commands exist: copy, insert, delete, and replace. Take for instance the transition from δοραιαν to δορεαν. Our algorithm, built on the

28 Elliott – Bodard – Cayless *et a*l. 2006.

29 See also Vierros – Henriksson 2017 and Vierros 2018.

30 E.g. Stolk 2018.

31 For a detailed discussion, see Gusfield 1997.

32 Ratcliff – Obershelp 1988.

difflib-based string comparison method, suggests the following instructions: copy δορ, replace αι with ε, and then copy αν. We also consider the surrounding environment of the variants; in this example, αι is preceded by δορ and followed by αν, a potentially interesting context for this spelling error. Cases with multiple errors within a single word, like γιτοονος corrected to γειτονες, are managed by our system through a combination of copy, insert, delete, and replace commands.

3.3 Database structure

As explained above, our system handles the differences between the original and regularized forms through a series of edit instructions. For effective retrieval, these edits and their adjacent context are stored in an indexed MySQL database table, with dedicated text fields for each one ("original," "regularized," "original_before," "regularized_before," "original_after," "regularized_after"), along with a field indicating the edit operation in question ("copy," "insert," "delete," "replace"). Furthermore, to enable case and diacritic-insensitive searches, the table includes fields for de-accented and lowercase versions of these strings.

The variation table further includes a "token_id" field that links each indexed variation to its corresponding word in the database. This arrangement not only allows querying instances where an editor has altered a nonstandard form, but also enables more complex queries that integrate the change with morphology, syntax, and document metadata. This extended functionality of PapyGreek Search is elaborated in Henriksson – Vierros, this volume; here we primarily utilized its variation search function.

3.4 Search Queries for Variations

The PapyGreek Search interface is available at https://papygreek.com/search, where users are presented with a main search box where search terms can be entered using specific parameters. For the most basic word searches, the user would input "form=" followed by the desired term (without diacritics), such as "form=και" to find all instances of και. For queries specifically targeting linguistic variations, a special syntax is implemented. Users apply the "form" parameter in combination with symbols representing editorial actions: + (plus) for additions, - (minus) for deletions. As an illustration, "form=-ι" would yield all instances where ι has been deleted by an editor, while "form=+ι" would reveal all cases where ι was added.

Finding editorial replacements requires a combination of the - and + symbols. If one wishes to find instances where ε has been replaced by η, the search term would be "form=-ε+η". To further refine the search based on context, the symbols > (before) and < (after) are used. For instance, to find instances where ε has been corrected to αι following λ, the query would be "form=λ>-ε+αι". Finally, PapyGreek Search handles re-

gular expressions through the “regex:form=” parameter, which allows for the creation of more intricate and specific search criteria.

4 Case studies

In this section, we will illustrate how PapyGreek Search can yield novel and potentially significant results for previously unexplored phonological phenomena. There are two especially interesting cases that involve Egyptian-Coptic phonological impact visible in the nonstandard Greek orthography that show unclear or conflicting results: (a) how the adjacency of /r/ affects vowel quality, and (b) how the Egyptian-Coptic-influenced /o, u/ allophonic variation in Greek texts gets confused with nonstandard case inflection, and is usually taken to only mean morphological variation. We will start with the phenomenon of /r/ both fronting and retracting vowel quality, all the while still following a phonological symmetry in doing so: the actual nature of this consonant seems very liquid indeed in coarticulation, often merely transferring information of phonemes near it to other phonemes around it.

4.1 Case study (a): vowel variation related to the adjacency of /r/

The phoneme /R/, which in sociolinguistic (or sociophonetic) studies represents all variants associated with /r/-like sounds, regardless of what they are precisely phonetically, can be understood as a wider representative for all the phoneme qualities associated with what is written with one letter only in the world’s languages: <r>. This is a phoneme with a particularly wide range of variation for a consonant. It seems that /r/ can be almost anything that is an oral lingual sonorant consonant that is not particularly palatal (such as /j/), lateral (such as /l/), or bilabial (such as /m, b, p/). While there is phonetically no natural class that forms ‘rhotics’, they are nevertheless considered as part of the same phonemic group by most language speakers. What exactly is ‘rhoticity’ is unclear among phoneticians, as for instance a Spanish trill can be pronounced as a tap by some speakers, and these taps can be produced differently from one speaker to another; nevertheless, there is an intuitive connection between the various sounds written with the grapheme <r>, and the variants mentioned above are understood as allophones of the one and only /r/ in the language.[33]

Despite this one grapheme <r> being the only letter describing the phoneme in written language, /R/ actually has quite a few representatives phonetically, not all belonging to the same category. There is an alveolar trill /r/, as is in use in e.g. Finnish and Italian;

33 Scobbie 2006, 338–9; see also Rennicke 2015 for (socio)phonetically different variants of /R/ in Brazilian Portuguese

there is an alveolar tap (sometimes also called a flap) /ɾ/, which can be heard in e.g. American English when pronouncing a geminate /t/ as in *better* [bɛɾəɹ]; there is a retroflex approximant /ɻ/ as can be heard in e.g. Indian English (and generally in Indian native languages); and there is an alveolar approximant [ɹ] as can be heard in British English rhotic dialects, as well as in American English, when pronouncing e.g the word-final /r/ in *better*. In some languages, /r/ pairs up with /l/, both of which are called liquids, and these are not understood as individual phonemes; such is the case in e.g. Japanese. As can be seen from all this, /R/ is phonetically interesting.

In Greek misspellings coming from the Roman-Byzantine period Egypt, there is variation worth noting related to whether the adjacent vowel quality to /r/ is fronted or retracted, which both occur. Even more than in Greek texts, this is noticeable in Greek loanwords in Coptic texts, which naturally show even more integration to native language phonology than the second language (L2) Greek used by what presumably was mostly Egyptian writers. This is because loanwords are most often treated as native language words and for this reason, they have undergone adaptation to the native language phonological system.[34] But why does /r/ affect both directions?

Phonetically, /r/ belongs to the coronal consonant group, consonants formed with the tip (apical) or blade (laminal) of the tongue.[35] (European) Coronal consonants are listed in Table 3 below.[36]

Table 3: (European) coronal consonants.

z	s	ð	θ	ʒ	ʃ	n	d	t	ɹ	l	r	ɾ

Because coronal consonants are formed with the front part of the tongue at the front (from dental to postalveolar) part of the oral cavity, they usually raise or front the nearby vowel's quality; only in the case of retroflex phonemes, in which the tongue tip literally flexes backwards, vowel quality can actually be retracted as well.[37] On the other hand, in many languages of the world, also other types of /R/-sounds seem to retract vowel quality; this is the case, for example, in Modern Arabic[38] and Swedish.[39] In Swe-

34 E.g. Haugen 1950, 215–22; Weinreich 1968, 26–7; Major 2001, 136–7.

35 Ladefoged – Maddieson 1996, 11.

36 According to Dixon 2002, Australian aboriginal languages contain some coronals that are not found in consonant inventories of the European languages. These include for instance nasal and lateral consonants with a dental articulation, as they contrast with so-called peripheral (i.e. non-coronal; labial and velar) consonants, which is a typical feature of these languages. In this study we have only included coronal consonants that are relevant for the Greek study.

37 Flemming 2003, 335–6.

38 E.g. Ryding 2005, 26.

39 Riad 2014, 18–21.

dish, the quality of the liquid is the actual retroflex approximant /ɻ/, but in Modern Arabic, it is the same alveolar trill as in e.g. Finnish.

The answer to this seemingly curious feature of the phoneme use also in Egyptian-Coptic lies in the phonetic feature of /R/ universally. It is a somewhat 'weak' phoneme that eventually even tends to disappear from languages through lenition,[40] and this is probably also the reason it seems to affect vowel quality in two opposite directions. /R/ is (literally) so liquid in nature that it can take coarticulatory effect even from the syllables before or after the immediately adjacent phoneme, in anticipatory or carryover coarticulatory effect. In the case of consonant-to-vowel coarticulation, i.e. the process of vowels adapting to the manner or place of the adjacent consonants, this means that the adjacent vowel's quality can likewise be altered to follow the quality of the phoneme before or after /R/, the liquid merely transferring these phonetic properties onto the vowel. In effect, what this means is that if there is a front phoneme, consonant or vowel, before or after /R/, the vowel affected by /R/ in coarticulation will be fronted. If there is a back phoneme before or after /R/, the vowel next to it will be retracted in quality.

From what can be gathered from the evidence of this phenomenon in Egyptian Greek is that it seems that being the weak phoneme /R/ is, regardless of its precise quality in Egyptian-Coptic (or Greek), it is indeed itself affected by phonemes in the syllables around it, and thereafter transfers the phonetic properties of them onto the vowel qualities adjacent to it. The easiest phoneme variation pair with which to show this is eta/epsilon (in the Postclassical era phonemically ε, η /e, ę/[41]) due to this being among the only ones that cannot be confused with, for example, case variation, which is a very common problem when studying the round vowel (ο, ου /o, u/) variation. In addition, this goes toward proving that eta had not raised to /i/ at this point of the Egyptian Greek internal phonological development, as often assumed;[42] we find evidence both for synchronic contact effect and diachronic phonological development at the same time, with the same search parameters. In phonetic quality, eta represented the near-close front unrounded vowel /ę/, but its quality seems to have been positionally variable.[43] However, as there are, naturally, also other examples of the same fronting/retracting phenomenon connected to the adjacency of Egyptian Greek /r/, and instances where this happens near the other liquid consonant, /l/,[44] we have also included /i/ and /y/ into the search: in the examples, there are also instances in which even /i/ and /e/ are in variation, and /y/, naturally, is a frequent phonetic variant of /i/.

Some examples of the results of the search are presented in Table 4 below, and Figure 1 shows the timeline of the variation: it mostly takes place in the centuries of the emer-

40 See Rennicke 2015 for discussion and analysis.

41 Horrocks 2010, 112.

42 Teodorsson 1977; Horrocks 2010, 165–70.

43 Horrocks 2010, 167–70; Dahlgren 2017, 103–6.

44 Dahlgren 2017, 100–6.

gence and strongest use of Coptic.[45] This seems natural: at least some of the Egyptian L2 Greek writers used the same alphabetic writing system also in their own language, so they were familiar with the specific connection between phonology and orthography, creating misspellings that were based on forms originating from the spoken language.

In PapyGreek Search, the parameters used to examine variation adjacent to /r/ include using the regex mode and allowing /r/ to potentially occur with any vowel quality. However, we have intentionally omitted ο /o/, ω (phonemically /o:/ before the Postclassical era) and ου /u/. This decision was made to avoid the significant number of variants likely resulting from the merging of genitive and dative cases in the Postclassical era, which end in *-u* and *-o*, respectively. Since some of these seemingly case-related variants can also be phonetically-induced, such a search can be conducted separately to allow for a lighter phonemic analysis within those particular results. For the sake of brevity in this chapter, we have simply excluded this analysis, partly because there is never any certainty without looking into each instance separately to be able to decipher whether the variant could be seen as something phonetic (with similar misspellings occurring also elsewhere in the text) or whether it likely was regarding the famous case merger (with few phono-orthographic misspellings, but with other morphosyntactic variation present in the text). We have also excluded α /a/ from the search to avoid numerous examples of word-final vowel reduction to *schwa*, a phenomenon unrelated to the effect of the liquid consonant in the word, as in τεσσαρες from the standard τέσσαρας (e.g., bgu.4.1051, 30 BC–14 AD).[46] All in all, the search yields 887 entries in which the vowel quality is changed after the occurrence of /r/.[47] Run in the opposite direction, there are 1.460 variants of the vowel quality being changed before /r/. From this, we could deduce that Egyptian-Coptic coarticulation might have been more anticipatory in nature than carryover, but obviously reaching such a conclusion would require more extensive research on that precise phenomenon.

Table 4: Consonant-to-vowel coarticulation on the vowel following /r/.

Nonstandard	Standard	Document	Date	Provenance
κατακεχωρεκα	κατακεχώρηκα	cpr.1.198	138 AD	Arsinoites or Herakleopolites
μιλιαρισια	μιλιαρήσια	sb.6.9140	601 AD	Arsinoites (?)
βεριδαριου	βερεδαρίου	p.lond.4.1383	708-710 AD	Aphrodites Kome (Antaiopolites)
διατρεψαι	διατρῖψαι	bgu.4.1208	27-26 BC	Busiris (Herakleopolites)

45 Dahlgren 2017, 28–34.

46 See Dahlgren – Leiwo 2020 for this Egyptian-Coptic -induced phenomenon; also more in Case Study (b).

47 This search query targets vowel variations (ε,η,ι,υ) immediately following the consonant ρ, using the syntax "regex:form=ρ$>+^[εηιυ]$-^[εηιυ]$". Data was retrieved on July 1, 2023.

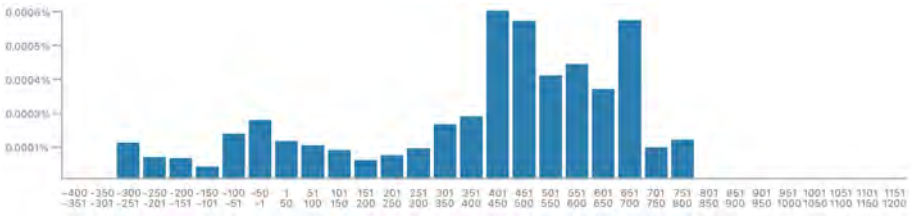


Fig. 1: Relative frequencies of words with vowel variation (ε, η, ι, υ) after ρ.[48]

In the first example shown in Table 4, κατακεχωρεκα from κατακεχώρηκα, the nonstandard production has an epsilon /e/ after /r/ instead of the standard eta /ę/. This is likely due to the retracting effect of /k/, a velar consonant, in the following syllable, which counts as one of the back consonants.[49] Μιλιαρισια from μιλιαρήσια, on the other hand, has a fronted nonstandard variant of the standard eta /ę/; it has raised to /i/. Given that the document dates to 601 AD, which is later than that of our first example, the change might be considered to reflect an advanced stage of completion of the raising of eta to /i/. However, there are many examples in Egyptian Greek documents of eta still being in variation with epsilon even in later centuries than this,[50] so it is equally likely that this is still phonetic variation. In general, the vowel quality in Greek used in Egypt was often retracted near /s/.[51] However, as the said /s/ is itself still followed by /i/ and the previous syllables contain two /i/'s, there is a very strong possibility of the combination having resulted in a raised vowel quality regarding eta, the liquid consonant just acting as a bridge between the front vowels.

Another, perhaps clearer and chronologically later example of a raised vowel quality adjacent to /r/ can be seen in the nonstandard variant βεριδαριου from the standard βερεδαρίου. Here, an /i/ replaces (epsilon) /e/. The vowel is positioned between /r/ and a front consonant, the coronal consonant /d/, which (together) evidently raise the vowel quality. Last, διατρεψαι from διατρῖψαι has a replacement of the standard word-medial iota with an epsilon, and in this example, the vowel quality seems to have been retracted by the combined forces of the /r/ before it and the bilabial (psi) after it. This earliest example, dating from the last pre-Christian century, should predate any poten-

48 Relative frequencies in the graph denote the proportion of word tokens from the search results in relation to the total number of word tokens in the entire papyri.info text database for each 50-year interval (e.g., 350–301 BC). Words from loosely dated documents spanning multiple 50-year periods were equally divided among these periods. For example, a document with 80 tokens covering a 200-year range allocates 20 tokens to each 50-year interval. For documents dated only before or after a specific time, the word count was distributed based on the average date range of 60 years, derived from fully dated documents in the database. Documents without any dating information were excluded from this calculation.

49 Jakobson 1968.

50 Dahlgren 2017.

51 Gignac 1976; see also Dahlgren 2017.

tial Coptic influence. In this particular instance, it is obviously difficult to say what role the liquid consonant may have had in the resulting variation as it could have been caused by the bilabial consonant alone (see more on the effect of the bilabials on vowels in the next section). Nevertheless, the examples above illuminate how varied results regarding a changed vowel quality can be found in the proximity of /r/.

4.2 Case study (b): vowel variation related to /o, u/ allophonic variation in Egyptian-Coptic

Greek documentary papyri have a substantial amount of variation between ο /o/ and ου /u/, which seems to be related to their allophonic status in (Egyptian-)Coptic.[52] Variation between these graphemes in Greek texts was already noted by Gignac,[53] who observed that this interchange often occurred in relation to stress: Greek unstressed ο /o/ was often replaced by ου /u/ and vice versa, which complies with the stress/allophonic rules of Coptic. Coptic had no unstressed ο /o/ (or /ɔ/) but it did have unstressed /u/,[54] and already Girgis remarked that Greek unstressed /o/ was often replaced with ου /u/ word-medially in Greek loanwords in Coptic.[55] In addition to this, this particular type of variation has been linked to the phonotactics of Coptic, i.e. the tendency of /o/ being replaced with /u/ in the adjacency of /m/ and /n/.[56] Although the coarticulation of consonants on vowels was a strong feature in Egyptian-Coptic, and the nasal/bilabial environments both have a tendency to raise the quality of open vowels crosslinguistically,[57] it is nevertheless very clear that this variation is not limited to these contexts. Native language prosody is typically one of the last elements to be lost by L2 speakers of a foreign language,[58] and it appears to be the case in L2 Greek writing in Egypt as well.

To confirm that this phenomenon is mainly related to stress and not coarticulation, there are some examples of variation that seem to indicate a change in the stress position. For instance, both λουγου /lugu/ in PSI VIII 884, 2 (390 AD) and κομιονται in BGU IV 1123, 6 (30 BC–14 AD) demonstrate variation related to the /o, u/ contact transfer. The standard forms of these words are λόγου /ˈlogu/ “word (gen.)” and κομιοῦνται “to take care of”, respectively.[59] As can be seen, in λουγου the first syllable’s stressed ο /o/ has been replaced with ου /u/, seemingly indicating that for the writer, this was the unstressed syllable; the genitive ending might have been learnt by heart due to its high

52 Dahlgren 2017, 83–4.
53 Gignac 1976, 211.
54 Peust 1999, 250–4.
55 Girgis 1966, 81–5.
56 Horrocks 2010, 112; Peust 1999, 238–40.
57 Beddor 1983, 2015; Flemming 2009, 82–4, 92.
58 Gut – Trouvain – Barry 2007; Matras 2009, 231–3.
59 See Dahlgren 2017, 153; in Postclassical Greek, the stress system had changed from having a primarily pitch accent to having dynamic word stress.

occurrence in Greek, being used in many patronymic forms. Similarly, in κομιονται the third syllable's original stressed ου /u/ has been replaced with ο /o/, as if to follow the Coptic phonemic distribution of /o/ being used as the rounded vowel in the stressed syllable, and /u/ in the unstressed one. This is probably due to transfer of Coptic stress rules, which fit in with the variant form's apparent word stress position.

Judging by the descriptions of other stress-timed languages, Coptic seems to have been one. It tended to place stress on one of the last two syllables of the word.[60] It also seems that, typically for stress-timed languages,[61] the stress was placed on the heavy syllable, at least in disyllabic words; perhaps more related to the word stress position, i.e. typically near the middle part of the word, in longer ones.[62] Variation, therefore, between ο, ου /o, u/ could be explained by (Egyptian-)Coptic stress rules and the phoneme distribution related to them. The vowel group α, ε, ο /a, e, o/ was also subject to neutralization in an unstressed position, especially word-finally.[63] /a, e, o/ variation concerned especially verb semantics, and could cause confusion over whether hybrid verb formations such as κερασεν (κέρασον) or πεμψεν (πέμψον) were to be interpreted as infinitives or imperatives.[64] Similarly, the stress position of the replaced vowel quality in κομιονται matches Coptic stress rules:

1. it is on one of the last two syllables;
2. it is in the middle part of the word, although four syllables can not be parted exactly in the middle;
3. it also happens to be the heavy syllable of the word with two consonants following the vowel.

On studying the phenomenon related to the /o, u/ variation in Egyptian Greek, we made four separate searches for the variation between ου/ω and ο/ου. The first two of these, standard ω replaced with ου, and vice versa, give many results that are usually interpreted to stem from the genitive and dative case merger, visible in the word-final variation of ου and ω (in Postclassical Greek, phonemically /u, o/); this was the position of Greek cases. The first search yields 1,348 results.[65] Searched the other way around there are 966 instances of standard ου replaced with nonstandard ω.[66] From the conso-

60 Loprieno 1995, 37; Peust 1999, 273.

61 Nübling – Schrambke 2004, 284–5.

62 Dahlgren 2017, 83–4, 153.

63 Dahlgren 2017, 62–6.

64 Dahlgren – Leiwo 2020.

65 The search targets variations where the original form has ου /u/ and the regularized form has ω /o/. The search query ("form=-ου+ω") was executed on 1st July 2023. Out of the 1,348 results, 1,161 are word final.

66 The data with the query ("form=-ω+ου") was retrieved on 1st July 2023. We note, however, that on some occasions the writers used (silent) *iota adscript* in connection with the ω, e.g. to mark the singular dative case (-ωι), and if we want the query to include these variants, we should use the regex query ("regex:form=-(ω|ωι)$+ου$"); this gives us 1,068 results. When we add a restriction of word final position ("regex:form=-(ω|ωι)$+ου$<^$"), we get 839 results, which means that only 229 instances are *not* word final.

nant environments, it is clear that some of the examples seem more like involving case merger, such as the nonstandard production δεισκου /deisku/.

In δεισκου from δείσκῳ, one of the many examples of what seem to be case confusion, the word-final stressed /o/ has been replaced with /u/, effectively changing a dative case to a genitive one. The nonstandard form also has a raised vowel quality from /o/ to /u/ after /k/, a velar consonant that, at least theoretically, should more retract the vowel quality, so there is no easy explanation to link it to coarticulation. The personal name misspelling Πετρου from what would have been the standard here, Πέτρῳ, on the other hand shows the possible effect of case merger, the position of stress and the related allophonic distribution of /o, u/ in Coptic, as well as consonant-to-vowel coarticulation: the standard /o/ could have been raised under the influence of the preceding /t/ and /r/, both coronal consonants. Πετρου is a prime example of how complicated it is to categorize variation in Egyptian Greek. Ορμου from ὅρμῳ looks like a clearer case of case merger as the preceding consonant is a bilabial /m/, although it primarily lowers high vowels instead of raising lower ones, which has happened here. Μαλλουṭον from μαλλωτὸν is interesting, especially from the point of view of the Greek stress position; below, we will talk about the possibility of transfer of stress, but in this example, the stress position seems to have been faithfully kept, and seems to be reflected in the replacement of the original ω /o/ to ου /u/, which is the unstressed rounded vowel in Coptic phonology. Consonant-to-vowel coarticulation is another possibility, of course, with the surrounding consonants /l/ and /t/ of the vowel being coronal ones. Also in αλλου /'allu/ from ἄλλῳ /'allo/, the word-final unstressed /o/ has been replaced with /u/ as per Coptic allophonic stress rules. In addition to the contact-induced stress connection, the coronal consonant /l/ could in this instance be raising the vowel quality. However, with the multitude of these types of cases altogether, stress-related variation seems a likely scenario for many, if not most, of the nonstandard vowel qualities because the word-final vowel, on which the case marking rests, is often unstressed in the Greek standard forms, especially in disyllabic words.

Some examples of the first search can be seen in Table 5, and again Figure 2 shows the distribution over the centuries – again highlighting a peak in the centuries when Coptic was used.

Table 5: Variation between ου and ω.

Nonstandard	Standard	Document	Date	Provenance
δεισκου	δείσκῳ	p.brem.24	116 AD	Hermopolis (?)
Πετρου	Πέτρῳ	p.brook.16	651-700 AD	Krokodilopolis (Arsinoites)
ορμου	ὅρμῳ	p.cair.isid.15	309-310 AD	Karanis (Arsinoites)
μαλλουṭον	μαλλωτὸν	p.cair.masp.1.67006v.	566-570 AD	Antinoopolis (?)
αλλου	ἄλλῳ	p.freib.2.8	144 AD	Unknown

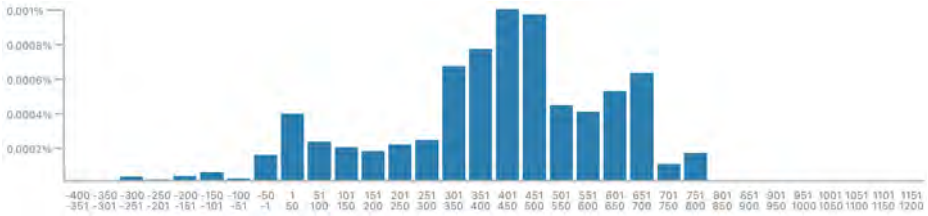


Fig. 2: Relative frequencies of instances of ου standardized to ω.

Table 6 shows the third search, related to the same variation of /u/ and /o/ but between the graphemes ου and ο, also proving that the quantity difference distinguishing *omicron* and *omega* had disappeared. Interestingly, the timeline in Figure 3 shows a dif-ferent distribution than in Figure 2 related to omega and ου: in Fig. 2, the variation is predominantly visible in the 3rd to 4th centuries AD, giving firmer evidence of the group belonging to the actual case merger category. In Fig. 3, there is variation both in the pre-Christian centuries as well as after; a high peak in the first century AD and a steady peak in the 3rd to 6th centuries AD. All examples show a possibility of stress transfer involvement, as well as a possible coarticulatory effect. The search yields 578 tokens.[67]

Table 6: Variation between ου and ο.

Nonstandard	Standard	Document	Date	Provenance
τουτου	τοῦτο	p.grenf.2.30	102 BC	Pathyris
εικουσι	εἴκοσι	p.ant.1.42	557 AD	Lenaiu (Antinoites)
μενουντος	μένοντός	basp.51.49	345 AD	Oxyrhynchos

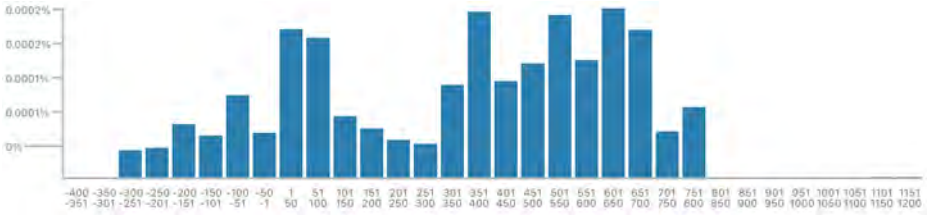


Fig. 3: Relative frequencies of instances of ου standardized to ο.

67 This search aims to find variants between ου /u/ and ο /o/ where the editor has corrected the nonstandard ου to ο. At the character-level, this corresponds to the editorial deletion of υ after ο, which in PapyGreek Search can be expressed as "regex:form=ο$>-^υ$". The search was conducted on 1st July 2023.

The fourth search, concerning the variation between omicron and ου but in another direction i.e. the standard ου being replaced with nonstandard ο, gives results that could, again, result from a number of things from case merger to coarticulation, but do also show a replacement of vowels that match Coptic stress rules. In μερος from μέρους, the replaced ο from the standard ου is on the second syllable, which is heavier, thus matching the stress position of the word that would have been more natural for Coptic, as do the next examples, αποδοναι, ησυχοντος, and ετομεν.[68] Figure 4 gives the distribution of tokens again, showing a peak in the later centuries from 4th to 6th, when Coptic was in use. The search gives 867 tokens.[69]

Table 7: Variation between *omicron* and ου.

Nonstandard	Standard	Document	Date	Provenance
μερος	μέρους	bgu.1.251	81 AD	Soknopaiu Nesos (Arsinoites)
αποδοναι	ἀποδοῦναι	bgu.2.595	75-85 AD	Arsinoites
ησυχοντος	ἡσυχοῦντος	sb.6.9138	576-600 AD	Arsinoites
ετομεν	αἰτοῦμεν	sb.6.9194	276-300 AD	Alexandria

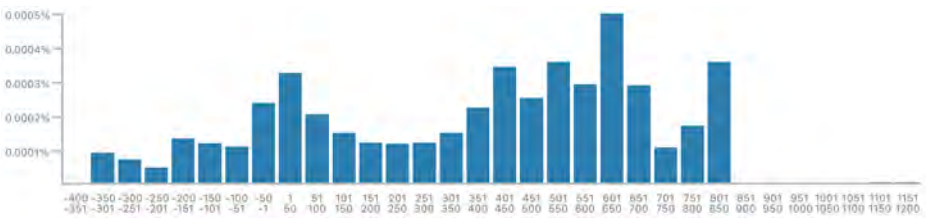


Fig. 4: Relative frequencies of instances of ο standardized to ου.

5 Discussion

In this chapter, we utilized the novel PapyGreek Search tool to examine phonological variation within Greek documentary papyri, focusing specifically on variations associated with the language contact situation between Greek and Egyptian-Coptic. We presented two case studies that explore potentially complex phonological variations. The

68 Dahlgren 2017, 133–8.

69 This search targets instances of nonstandard ο /o/ in place of the standard ου /u/; in other words, editorial additions of υ after ο. The search ("regex:form=ο$>+^υ$") was executed on 1st July 2023.

types of variation we chose for the study are notably challenging to interpret from textual evidence, in comparison to, for example, exploring more straightforward phonetic coarticulation that can be directly seen in the misspellings.

The first case study examined the variation of vowel qualities adjacent to /r/ in Greek documentary papyri. Through the analysis of specific examples, the study provided evidence suggesting that liquids may function as phonetic bridges in Egyptian Greek, transferring the phonetic properties of neighboring phonemes to the vowel qualities near /r/. The observations discussed in this study were obtained using the regex feature of PapyGreek Search. This feature allowed for the targeting of multiple vowel variations using a single search query, a task that was not achievable using traditional methods or previous digital tools. However, it was also acknowledged that determining the exact causes of these variations is often challenging, as writing errors can stem from various factors including, but not limited to, coarticulation, language contact, or phonological changes. Moreover, we focused on individual examples, and future research would benefit from a larger sample size and statistical analysis to enhance the credibility of the analysis. Nonetheless, the study underscored the capabilities of PapyGreek Search, paving the way for further investigation into the phonological characteristics of Egyptian Greek, and through contact-induced transfer effects, also Coptic phonology.

Second, we scrutinized vowel variation in cases, a subject that still remains inconclusive and multifunctional.[70] Still, phonetic and phonological factors must not be discounted when they coexist with morphological variation. Notably, the variation between /o/ and /u/ was relatively rare before the Roman period — precisely when Coptic began to make an appearance. Few instances were recorded during the Ptolemaic period, with the real influx of this variation starting in the Roman period, evident in Greek texts and in Coptic renderings of Greek loanwords.[71] This suggests a potential link to (Egyptian-)Coptic phonological influence in some sort of a bilingual milieu, whether this be related to the spoken level or orthographic practices. For L2 Greek speakers, the distinction between vowel qualities may not have been audibly discernible, but there might be a learned practice on the level of orthography to use omicron only for the stressed rounded vowel quality. This, however, is something we will not be able to completely verify within text linguistics because it rests on evidence from actual spoken language, which remains beyond our reach. Nevertheless, we believe our case study illustrates that certain traces of the spoken language can be inferred, even when dealing with such complex phenomena as stress transfer. With the aid of a more extensive sample size, diverse search parameters, and statistical analysis facilitated by PapyGreek Search, Coptic stress patterns could possibly be exhaustively extracted from the L2 Greek data from Egypt.

70 See e.g. Stolk 2015.

71 Gignac 1976, 207 n. 2.

The main value of PapyGreek Search, for the purposes of the present study, was to serve as a fresh interface to the already existing data found in XML-encoded source files, specifically pairs of irregular word forms and their editorial corrections. The construction of this novel dataset involved using an algorithm designed to discern the character-level differences between the original and regularized word forms, with the identified "edit instructions" being stored in a MySQL database for efficient retrieval. This is not the first interface designed for the purpose of finding text irregularities in documentary Greek.[72] However, we believe that having several similar tools is beneficial to the field of digital papyrology. Used in conjunction, they can either validate results or raise questions on the findings, encouraging careful scrutiny and planning of data collection and search parameters.

Looking ahead, the variation search functionality in PapyGreek Search could be advanced in several ways. Firstly, our database currently includes only documentary papyri, representing just one of the openly available collections where linguistic variation is encoded in a machine-readable format. By expanding the PapyGreek Database to include other collections encoded in EpiDoc XML – such as literary papyri and many epigraphic documents – the search tool could offer a broader perspective on linguistic variation across different text types and linguistic registers. Secondly, it would be highly valuable if the tool could detect not only those irregularities that have been corrected and encoded in the texts but also the numerous non-standard forms that have gone unnoticed by editors.[73] One approach to achieve this could involve training an ancient Greek language model to identify and correct nonstandard spellings and grammatical forms. This could uncover a vast number of previously unknown textual irregularities and potentially significantly enhance our understanding of how Greek evolved during the Greco-Roman period.

Bibliography

Aho, A. V. (1991), *Algorithms for Finding Patterns in Strings*, in *Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity*, ed. by J. Van Leeuwen, Amsterdam, 257–300.

Beddor, P. S. (1983), *Phonological and Phonetic Effects of Nasalization on Vowel Height*, PhD Diss., Bloomington (IL).

Beddor, P. S. (2015), *The Relation Between Language Users' Perception and Production Repertoires*, in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1041.pdf.

Bellem, A. (2007), *Towards a Comparative Typology of Emphatics*, PhD Diss., London.

Consani, C. (1993). *La koiné et les dialectes grecs dans la documentation linguistique et la réflexion métalinguistique des premiers siècles de notre ère*, in *La koiné grecque antique. I Une langue introuvable?*, Nancy, 23–39.

72 Cf. Depauw – Stolk 2015.

73 Stolk 2018.

Dahlgren, S. (2017), *Outcome of Long-Term Language Contact: Transfer of Egyptian Phonological Features onto Greek in Graeco-Roman Egypt*, PhD Diss., Helsinki.
Dahlgren, S. (2020), *The System of Coptic Vowel Reduction: Evidence from L2 Greek Usage*, Italian Journal of Linguistics 32(1), 211–27.
Dahlgren, S. (2022), *Egyptian Greek: A Contact Variety*, in *Ancient Indo-European Languages between Linguistics and Philology*, ed. by M. Bianconi – M. Capano – D. Romagno – F. Rovai, Leiden, 115–52.
Dahlgren, S. – Leiwo, M. (2020), *Confusion of Mood or Phoneme? The Impact of L1 Phonology on Verb Semantics*, in *Postclassical Greek: Contemporary Approaches to Philology and Linguistics*, ed. by D. Rafiyenko – I. Serẑant, Berlin – Boston, 283–301.
Depauw, M. – Stolk, J. (2015), *Linguistic Variation in Greek Papyri: Towards a New Tool for Quantitative Study*, GRBS 55, 196–220.
Dickey, E. (2011), *The Greek and Latin Languages in the Papyri*, in *The Oxford Handbook of Papyrology*, ed. by R. Bagnall, Oxford, 149–69.
Dixon, R. M. W. (2002). *Australian Languages: Their Nature and Development*, Cambridge.
Elliott, T. – Bodard G. – Cayless, H. *et al*. (2006). *EpiDoc: Epigraphic Documents in TEI XML*, https://epidoc.stoa.org.
Flemming, E. (2003), *The Relationship Between Coronal Place and Vowel Backness*, Phonology 20(3), 335–73.
Gignac, F. T. (1976), *A Grammar of the Greek Papyri of the Roman and Byzantine Periods*, I (*Phonology*), Milan.
Gignac, F. T. (1991), *Phonology of the Greek of Egypt*, *Influence of Coptic On the*, in *The Coptic Encyclopedia*, ed. by A. Suryal Atiya, New York, VIII, 186–8.
Girgis, W. A. (1966), *Greek Loan Words in Coptic (Part II)*, Bulletin de la Société d'Archéologie Copte 18, 71–96.
Gusfield, D. (1997), *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, New York.
Gut, U. – Trouvain, J. – Barry, W. J. (2007), *Bridging Research on Phonetic Descriptions with Knowledge from Teaching Practice: The Case of Prosody in Non-Native Speech*, in *Non-Native Prosody. Phonetic Description and Teaching Practice*, ed. by J. Trouvain – U. Gut, Berlin, 3–21.
Hardcastle, W. J. – Hewlett, N., (2000), eds., *Coarticulation: Theory, Data, and Techniques*, Cambridge.
Haugen, E. (1950), *The Analysis of Linguistic Borrowing*, Language 26(2), 210–31.
Horrocks, G. (2010), *Greek: A History of the Language and Its Speakers*, 2nd ed., Malden (MA).
Jakobson, R. (1968), *Child Language, Aphasia and Phonological Universals*, The Hague [1941].
Ladefoged, P. – Maddieson, I. (1996), *The Sounds of the World's Languages*, Oxford.
Layton, B. (2000), *A Coptic Grammar*, Wiesbaden.
Loprieno, A. (1995), *Ancient Egyptian. A Linguistic Introduction*, Cambridge.
Major, R. C. (2001), *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology*, Mahwah.
Manuel, S. (1999), *Cross-Language Studies: Relating Language-Particular Coarticulation Patterns to Other Language-Particular Facts*, in *Coarticulation. Theory, Data and Techniques*, ed. by W. J. Hardcastle – N. Hewlett, Cambridge, 179–98.
Matras, Y. (2009), *Language Contact*, Cambridge.
Mayser, E. – Schmoll, H. (1970), *Grammatik Der Griechischen Papyri Aus Der Ptolemäerzeit. Laut- Und Wortlehre: Einleitung Und Lautlehre*, Berlin.
Nübling, D. – Schrambke, R. (2004), *Silben- versus akzentsprachliche Züge in germanischen Sprachen und im Alemannischen*, in *Alemannisch im Sprachvergleich. Beiträge zur 14. Arbeitstagung für alemannische Dialektologie in Männedorf (Zürich) vom 16. – 18.9.2002*, ed. by E. Glaser – P. Ott – R. Schwarzenbach, Stuttgart, 281–320.
Peust, C. (1999), *Egyptian Phonology: An Introduction to the Phonology of a Dead Language*, Göttingen.
Ratcliff, J. W. – Metzener, D (1988), *Pattern Matching: The Gestalt Approach*, Dr. Dobb's Journal 46.
Reggiani, N. (2017), *Digital Papyrology I. Methods, Tools and Trends*, Berlin – Boston.
Rennicke, I. (2015), *Variation and Change in the Rhotics of Brazilian Portuguese*, PhD Diss., Helsinki.
Riad, T. (2014), *The Phonology of Swedish*, Oxford.

Ryding, K. C. (2005). *A Reference Grammar of Modern Standard Arabic*, Cambridge.
Scobbie, J. (2006), *(R) as a Variable*, in *Encyclopedia of Language and Linguistics*, X, 337–44.
Stolk, J. (2015), *Case Variation in Greek Papyri: Retracing Dative Case Syncretism in the Language of the Greek Documentary Papyri and Ostraca from Egypt (300 BCE–800 CE)*, PhD Diss., Oslo.
Stolk, J. (2018), *Encoding Linguistic Variation in Greek Documentary Papyri: The Past, Present and Future of Editorial Regularization*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 119–37.
Teodorsson, S.-T. (1977), *The Phonology of Ptolemaic Koine*, Gothenburg.
Thomason, S. G. – Kaufman, T. (1988), *Language Contact, Creolization, and Genetic Linguistics*, Berkeley (CA).
Torallas Tovar, S. (2010), *Greek in Egypt*, in *A Companion to the Ancient Greek Language*, ed. by E. Bakker, Malden (MA) – Oxford – Chichester, 253–66.
Vierros, M. (2012), *Bilingual Notaries in Hellenistic Egypt. A Study of Greek as a Second Language*, Brussels.
Vierros, M. (2018), *Linguistic Annotation of the Digital Papyrological Corpus: Sematia*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 105–18.
Vierros, M. – Henriksson, E. (2017), *Preprocessing Greek Papyri for Linguistic Annotation*, in *Journal of Data Mining and Digital Humanities. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*, ed. by M. Büchler – L. Mellerin, http://jdmdh.episciences.org/paper/view/id/1385.
Weinreich, U. (1968), *Languages in Contact: Findings and Problems*, The Hague [1953].

Part III: **Materiality, Textuality, and Scribal Phenomenology**

Daniel Riaño Rufilanchas

Callimachus: A Digital Regest of Greek and Latin Papyri

1 Introduction

In the 21st century, the most common form of circulation of papyri has undoubtedly been electronic, probably in some type of XML format. Among the various standards for encoding epigraphic and papyrological texts, the most successful to date is the TEI subset known as EpiDoc.[1] Initially designed for epigraphic texts, this schema has proven equally adaptable to papyri, which share the characteristic of generally being individual texts, often fragmentary, that usually have an initial edition and subsequent editions or corrections in modern times and of which several copies or total or partial reformulations may survive (unlike, for example, classical texts that depend on the collation of several manuscripts).[2]

Naturally, as in any copying process, the reader (in this case, the user of the digital document) loses much of the information directly transmitted through the material qualities of the original document, such as the quality of the support or the nuances that a contemporary speaker of the author might perceive in subtleties affecting the layout of the text, the tracing of the characters, etc. Despite this, the XML format is undoubtedly more useful than traditional paper editions (or the collation of the original document itself) for the automatic or semi-automatic processing of large quantities of papyri necessary for a multitude of tasks related to the papyrological, philological, historical, or linguistic study of papyri. Moreover, a significant portion of this information about the document itself (its internal layout, its material characteristics) and its history (its origin and details about its current whereabouts or the editing process) can be encoded as metadata or markup elements.

A large part of the digitally edited papyri to date has a copy accessible thanks to the cooperative project Papyri.info.[3] As of April 2024, this project included nearly 69,000 documentary papyri (DDbDP) and 14,800 literary or paraliterary papyri (DCLP) as XML files encoded in EpiDoc, along with metadata corresponding to a large portion of them from HGV and APIS.

1 See Bodard 2010.

2 Literary papyri can naturally result from an ancient collation of previous texts, but this does not prevent us from considering, in general, that the editing of epigraphic and papyrological texts is a process with different requirements from the editing of classical works, which usually demand a critical apparatus and a prior stemma.

3 A comprehensive review of the project, complete for the year it appeared, can be found in Vannini 2010.

https://doi.org/10.1515/9783111070162-013

To effectively utilize such a vast corpus of material for academic purposes, it is essential to develop digital tools capable of processing both the text and metadata, catering to diverse research interests. Three ways to utilize metadata can be: a) conducting searches on such metadata;[4] b) creating lists of the overall results obtained from the comprehensive examination of the documents; and c) building new databases from the processing of the obtained data. The most notable example of the latter is Trismegistos, which combines Papyri.info and other data sources to create prosopographic and toponymic databases, or Text Irregularities, a database of orthographical variations in Greek papyri, which has become one important tool for analysis of Greek phonetics and phonology.[5]

Researchers working on papyri often focus on their textual content, engaging in tasks such as tokenization, lemmatization, POS-tagging, and syntactic and semantic analysis from a linguistic perspective. Additionally, the material features of papyri have been meticulously studied by various groups, who have defined and enhanced the characteristics of EpiDoc across successive versions of this annotation standard. Callimachus is a tool specifically designed for searching and analyzing the material and non-linguistic data of papyri.[6]

2 The ecosystem of Papyri.info

Papyri.info is an international cooperative project for the creation and maintenance of a digital corpus of papyri (initially Greek, Latin, and Coptic) and the resulting aggregator of papyrological data.[7] Structurally, it consists of two parts: the Papyrological Navigator, which "supports searching, browsing, and aggregation of ancient papyrological documents and related materials," and the Papyrological Editor, which "enables multi-author, version controlled, peer reviewed scholarly curation of papyrological texts, translations, commentary, scholarly metadata, institutional catalog records, bibliography, and images."[8]

4 A part of such data can be searched using Papyri.info's own search engine at https://papyri.info/search. All hyperlinks last accessed on 1.6.2024.

5 https://www.trismegistos.org/textirregularities. See Stolk 2018; Reggiani 2019, 132–71.

6 The textual information of the papyri is handled by other projects, such as Polyphemus (https://glg.csic.es/Polyphemus/Polyphemus_presentation.html), also developed by the Greek Linguistics Group at ILC, CSIC.

7 Documentation on each specific project is not always extensive or easy to find. Probably the place that gathers the most information about most of the projects we cite in this section is https://papyri.info/docs/resources and the wiki page of the Digital Classicist https://wiki.digitalclassicist. org/Main_Page.

8 Quotations from the homepage https://papyri.info.

From the content perspective,[9] on one hand, Papyri.info consists of the documentary papyri gathered by the Duke Databank of Documentary Papyri (DDbDP)[10] co-directed by Joshua D. Sosin (Duke University) and James Cowey (Universität Heidelberg) and the literary papyri gathered by the Digital Corpus of Literary Papyri (DCLP), a project of the Institut für Papyrologie (Universität Heidelberg) and the Institute for the Study of the Ancient World (New York University) co-directed by Roger Bagnall and Rodney Ast.[11] On the other hand, regarding the metadata of such papyri, it integrates material from the Advanced Papyrological Information System (APIS),[12] the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV),[13] the Bibliographie Papyrologique (BP), co-directed by Alain Martin, Alain Delattre, Paul Heilporn, and Naïm Vanthieghem,[14] and more recently, the Arabic Papyrological Database (APD).[15] As a fundamental piece to allow the mapping and identification of material from each collection, a stable identifier, the Trismegistos number,[16] is used. In the rest of the section and the chapter, I will limit myself to what concerns Greek and Latin papyri, with only occasional references to other papyri such as Egyptian, Coptic, or Arabic.

Papyri.info is one of the most notable examples of successful academic collaboration in the field of Digital Humanities. This can be verified in three aspects. The first is the way its use has become widespread within its academic field: most of the ongoing papyri editing projects contribute to Papyri.info. The second is the quality of the database content, thanks to the operating protocol of the Papyrological Editor, which greatly facilitates the collaboration of external teams without subjecting them to excessive formal restrictions while ensuring high philological standards and the updating of the database with research following the publication of printed texts.[17] The third aspect is the very reach of this collaborative work's result, as virtually the entire community of papyrologists uses Papyri.info almost daily in their work, as do linguists and historians primarily concerned with papyrological material.[18]

Contributing to this result, in addition to factors such as the robustness of the community that maintains the ecosystem, are a series of judicious decisions made in the project's initial stages. Among them is the adoption of EpiDoc, a subset of TEI adapted for the annotation of ancient inscriptions and papyri, which is today a *de facto* standard

9 https://papyri.info/docs/about.

10 https://papyri.info/docs/ddbdp.

11 https://gepris.dfg.de/gepris/projekt/236701214?context=projekt&task=showDetail&id=236701214&>.

12 https://web.archive.org/web/20121222011708/http://www.columbia.edu/cu/lweb/projects/digital/apis/about.html.

13 https://aquila.zaw.uni-heidelberg.de/start.

14 http://www.aere-egke.be/BP.

15 https://www.apd.gwi.uni-muenchen.de/apd/project.jsp.

16 https://www.trismegistos.org/about_how_to_cite.php.

17 The page https://papyri.info/ has nearly 2,000 citations on Google Scholar. [June 1, 2024]

18 For an extensive description of the project, see Vannini 2010; Reggiani 2019, 50–5.

as a coding system for papyrological documents (and likely soon in epigraphy). According to the EpiDoc guidelines, "EpiDoc addresses not only the transcription and editorial preparation of the texts themselves, but also the history, materiality, and metadata of the objects on which the texts appear."[19] This means that searches on these documents can concern both the text's content and the other material aspects that have been tagged in XML or appear in the metadata.

In order to understand how projects in a 'second ring' of this ecosystem can operate (i.e., those like Callimachus that provide the community with some value derived from the integration of data and resources from Papyri.info itself, with the possible addition of other resources), it is important to consider that Papyri.info is a system that fundamentally grows through the voluntary collaboration of individuals and institutions that add their resources to the initial integration effort of the project. This implies that it is not expected for each collaborator to strictly adhere to a rigid annotation standard, a strictly limited repertoire of tags, or the same (natural) language to describe the texts. The project leaders have preferred to encourage and stimulate the collaboration of a significant number of people rather than impose very strict criteria, which has allowed the joint project to achieve its current dimensions and impact. Projects like Callimachus adapt to this circumstance.

As a final observation in this regard, I note that the type of metadata in DDbDP and DCLP does not coincide, and for many papyri, data on material aspects such as form, color, size, etc., are missing. Some collections use elements or attributes that respond to very specific annotation needs, such as the `<seg>` tag with `@type="other Layer"` to mark the *sovrapposti* or *sottoposti* of the Herculaneum papyri.

Regarding the DDbDP documents, 60,034 (87.2%) contain the edited text of the papyri, while for 8,824 of them (12.8%), only the metadata have been incorporated so far. The corresponding figures for DCLP, where the proportions are reversed, are 1,944 (13.1%) and 12,841 (86.9%).

There is metadata information from HGV for 68,615 DDbDP papyri (99.6%; this information is missing for only 243 of them, 0.4%). Since the definition of 'document' is not the same for both projects, there is not always an individual correspondence between the files in both databases, with it being relatively common for two or more HGV documents to correspond to the same DDbDP document. In the case of literary papyri, the percentages are again reversed, as we only have HGV metadata for three DCLP papyri. Regarding APIS, there is material in this database for 6,996 (10.2%) DDbDP documents.

19 https://epidoc.stoa.org/gl/latest.

3 What is Callimachus

Callimachus is an online database of Greek, Latin, and Coptic papyri built from the text, metadata, and markup elements of texts from DDbDP and DCLP, and the information about the same papyri gathered by the HGV and APIS projects.[20] It focuses on the non-lexical information of the papyri: that is, its categories consist of data on the materiality of the medium and its history (dating, material, provenance, location, etc.) and the material, non-linguistic elements of the text, such as the number of words, number of letters per line, the state of preservation and readability of the text, the type of script, the number or type of corrections, etc. It can therefore be considered a special type of digital regest of Greco-Roman papyri from DDbDP and DCLP. Unlike a normal regest, Callimachus does not yet offer a summary of the document's text content, but will soon do so via an AI-generated summary. The purpose of Callimachus is to offer a comprehensive tool that allows for:

a) Querying the occurrence of any material feature or combination of features in the papyri, as encoded in the document markup or recorded in the HGV or APIS databases.
b) Directing users to the specific papyrus where the identified features are located.
c) Providing aggregate statistics on the corpus of papyri for various purposes. For instance, generating statistics on the number of words in papyri from each century to support any statistical study on the evolution of linguistic phenomena traceable in the papyri.
d) Offering a standardized metric (the Callimachus Number) to assess the state of preservation and readability of each document.

Currently, Callimachus can be used primarily in two ways: performing sear-ches through the website, or utilizing the data summaries published on the Greek Linguistics Group's page, such as "Counting the number of words in Greek and Latin Papyri".[21] The data will be accessible by the end of 2024 in a public repository.[22] Callimachus can be used in any work that requires knowledge of the formal characteristics of a specific papyrus or any set of papyri. This makes it particularly suitable for certain tasks, such as research on the materiality of written culture in Egypt, the construction of textual corpora and treebanks, etc.

20 https://glg.csic.es/Callimachus/Callimachus_search.html.
21 https://glg.csic.es.
22 https://github.com/danielrruf/Callimachus.

4 How Callimachus is constructed

Each new 'version' of Callimachus is built by processing the latest version of the papyri from DDbDP and DCLP.[23] Each text is reprocessed, and the documents added since the previous version retain a stable numbering. The result of this processing is made accessible to all users at least twice a year to incorporate the ongoing contributions to Papyri.info. This process is done in parallel for documentary and literary papyri, using primarily Python and LiveCode scripts, drawing data from the project's main Github repository.[24] Callimachus is built in parallel with Polyphemus, a database of the lexical information of the papyri.

In the initial phase of building Callimachus, all metadata information related to the materiality of the document is extracted, primarily from the header of the EpiDoc document,[25] as well as from the HGV and APIS databases. Subsequently, the structural and markup elements related to the text (found in the `<body>` of the document) and the text itself are processed. The script records the occurrence of each attribute and element (such as `<milestone>`, `<g>`, `<gap>`, etc.) along with their attributes. It then re-tags the text so that each word includes all the XML tags affecting it, resulting in a string where each word is annotated with relevant tags, so that a string:

```
<lb n="35"/>ἀπ
<lb n="36" break="no"/>αλλ<supplied reason= "lost">αγῶμεν δοῦνα</sup-
plied><unclear>ι</unclear>
```

is re-tagged in this way:

```
<lb n="35-36"/>ἀπαλλ<supplied reason="lost">αγῶμεν</supplied>
<lb n="36"/><supplied reason="lost">δοῦνα</supplied><unclear>ι </unclear>
```

In the current state of DDbDP and DCLP EpiDoc documents, words are not tagged. Tokenization (determining which sequences of characters constitute a word) is not a trivial process when dealing with papyri. After combining those words that are divided between two (sometimes more than two) lines, it must be determined whether they are whole words, pieces of a word, or fragments of letters. The most frequent problem is the lack of insertion in the EpiDoc document of the space that separates words in modern editions before or after a tag. In addition to the inevitable errors in text entry, sequences of tags that prevent the correct insertion of spaces between words are very common.

23 The current version corresponds to the state of the Papyri.info Github repository as of 02.02.2024.

24 https://github.com/papyri/idp.data.

25 The structure of an EpiDoc document can be seen here: https://epidoc.stoa.org/gl/latest/supp-structure.html; the latest version of the schema can be seen at https://epidoc.stoa.org/schema/latest; a list of the EpiDoc Guidelines can be found at https://epidoc.stoa.org/gl/latest/app-alltrans.html.

Examples of such sequences include:

```
<unclear>τ</unclear><supplied reason="lost">ῇ<expan><ex>αὐτῇ</ex> </ex-
pan>·</supplied>τῇ<expan>ἀδελ<ex>φῇ</ex></expan>
μεθ’<hi rend="diaeresis">ὑ</hi>π<unclear>ο</unclear><supplied
reason="lost" cert="low">γραφῆς
```

This encoding may result in incorrect tokenization (e.g., τῇαὐτῇ, τῇἀδελφῇ, μεθ’ὑπογραφῆς) unless corrected. In these cases, Callimachus utilizes the Madrid Ancient Greek Word List (MAGWL), which contains more than two million forms and over 250,000 lemmas (including proper names). By leveraging MAGWL to identify pre-lemmatized forms and employing an algorithm to determine in ambiguous cases whether a sequence of letters constitutes one or multiple words – or if the editor considers it part of a word (e.g., the use of breathings and accents indicating the start of a word) – Callimachus generally achieves complete tokenization of a papyrus. Built in parallel with Polyphemus, Callimachus then lemmatizes and morphologically analyzes the identified forms using MAGWL. To disambiguate between possible morphological analyses of a form, we use UDPipe, provided that this resource returns a form among those selected as possible by MAGWL.[26]

When preparing a text for creating a treebank, it is crucial to consistently de-termine whether forms such as ὅταν should be considered as one word or two (ὅτε ἄν). Additionally, it is essential to treat results of crasis, such as τοὐπιόν from τὸ ἐπιόν (P.Oxy. LXXXIII 5362, 4) or κἀποσταλήτωσαν from καὶ ἀποσταλήτωσαν (P.Sorb. III 136, 9–10), as two distinct words. These types of linguistic considerations, however, are not of primary importance for the purposes of Callimachus.[27] It is important, for statistics on lexicon

26 https://lindat.mff.cuni.cz/services/udpipe.

27 I have counted 930 examples of crasis in the more than 4,600,000 words of the documentary papyri. The vast majority of these are reduced to a very short series of combinations. With more than 5 examples, we have: κἀγώ (228, compare with the 11 examples of καἰγώ), κἄν (221), τἆλλα (72), κἀμοί (50), κἀμέ (48), κἀμοῦ (22), τἄλλα (19), καὐτός (19), τἀντίγραφα (17), τοὔνομα (16), τἀντίγραφον (12), κἀκεῖνος (9), τἀργύριον (9), κούκ (8), καὐτοί (7), τἀναντία (7), κοὐ (6). The results are clearly of interest from a linguistic point of view, but they do not justify a change in the way words are counted for these statistical purposes (they are relevant for a project like Polyphemus, where each of the aforementioned compounds can be searched based on the elements that compose it). Oddly enough, most encoders have chosen to use the `<choice>` element to tag crasis, leaving the "main" element of the crasis as the only element of the regularized form and tagging the crasis as the original reading, as follows: `<choice><reg>ἀργύριον</reg><orig>τἀργύριον</orig></choice>`. This is an unexpected way to use this tag for text segmentation. A worse case occurs when the same element is used in a completely different manner, leading to examples like the following: `τὸ <choice><reg>ἀργύριον</reg><orig>ταργύριον</orig></choice>`. These cases have not been counted in the previous list because, in fact, the crasis does not appear in the XML document. On Greek and Latin treebanks, see Riaño Rufilanchas forthcoming.

and document types, to differentiate numerals (very numerous in many papyri) from the rest of the words.

To perform operations such as calculating letters per line, or calculating the Callimachus Number, it is also necessary to record whether the word is written in full, abbreviated, or represented using sigla. For forms corrected or regularized by the editor and tagged with `<choice>`, `<app>`, both the original and the edited forms are recorded separately.

The dating of a papyrus presents a wide range of cases: some documents are dated to an exact day of the month of a year that is relatively easy to assign to a date in our Gregorian calendar, and the date may (or may not) have been preserved in whole or in part. It is also possible that there is a reference to a specific event or person that serves as a reference point to place the writing at a specific date, or a specific year, or a month or day of a known or unknown year. Often, the archaeological context of the papyrological find is completely unknown, so the papyrologist must resort to internal criteria for dating within ranges that can be quite broad. The type of magistracies can indicate the period, the products mentioned can indi-cate the season, etc. Additionally, a papyrus may contain parts written at different times, leading to different ways of treating what is considered a 'document.' Finally, the interest in precision in the dating of a document can vary greatly from one project to another, or even within phases of the same project.

In addition to the issues briefly mentioned, it should be noted that there is no single way to annotate a specific date or time span. Forms such as 142/141 BC; 142 - 141 BC; 142 BC/ 141 BC; 142 - 141 BC; 142 or 141 BC; BC 142/141; BC142-141, 142-1 BC, BC 142?, etc., can all be found to refer to the same period of time. Adding to this variety are dates that can refer to two different periods or correspond to two phases of writing the same papyrus, expressed using different conventions, such as AD 341-374, 381-397?, or 217-6-200-199, etc. Often, uncertainties about the dating are expressed using "circa," or by referring to a century or a time window of several centuries (expressed with Arabic or Roman numerals), or to historical periods expressed by designations such as "Biz.," "Roman," etc. Such designations, common a few years ago, have almost completely disappeared, often replaced by the equally conventional form of 30 BC-AD 323 or even 30 BC-642 AD. Callimachus attempts to regularize these conventions and provide a way to perform searches within the period indicated by the user.

5 The Callimachus Number (CN)

The Callimachus Number (CN) is an algorithmic method to express the conditions of preservation and readability of the text preserved in a document or any portion thereof (a passage, a line, a word, etc.). 'Preservation' refers to the material state of the surface on which the text was originally written. 'Readability' refers to the possibility of recovering the original message from such a document or fragment, using not only the document itself but any other means available to the editor, such as context, the existence of

copies, parallels, citations, etc. In textual criticism, the difference between the preserved and readable (or reconstructible) text is expressed through specific conventions, such as the use of text within brackets in the Leiden convention. Typically, the readability index will be equal to or higher than the preservation index.

The CN is obtained by assigning a value between 0 and 1 to each linguistically valuable character in a document (each letter in the case of alphabetic scripts; each syllable in the case of syllabograms, etc.). The CN of any text is the sum of the CN of each character of the original text, divided by the total number of characters.

Three factors are considered when evaluating the Callimachus Number (CN): A) The degree to which a character is visible or recognizable; B) The presence of copies or texts that facilitate the reconstruction of lost or doubtful text with varying levels of certainty; C) The 'textual space' in which the illegible text or lacuna is situated, which helps estimate the extent of the lost text with greater or lesser precision. The 'textual space' can refer to the immediate context (the presence or absence of nearby characters) or be inferred from the document's layout, such as proximity to the edges of the page or column, and the presence of scribal markers indicating different parts of the discourse or text.[28] These three factors are applied progressively, following this algorithm (a Greek letter is used to locate each situation in the table). Note that the Callimachus Number (CN) may differ when measuring readability (RCN) and preservation (PCN), as there are instances where better preservation of a document section does not enhance message recovery, or conversely, the message can be recovered despite poor preservation due to other factors:

1. Characters are categorized into four distinct visibility situations: clearly visible, merely recognizable, unrecognizable ('traces'), or disappeared. In the first scenario (α), if the character is part of a recognizable word, it is assigned a Callimachus Number (CN) of one. For the remaining scenarios, the CN is determined by the subsequent criteria.
2. When a character is not clearly visible, the degree to which it can be determined that the character or its position in the document is part of a specific word is assessed. Four degrees of visibility are distinguished: A) A partially preserved character can be identified unambiguously with reasonable certainty due to the lexical and discursive context of the surrounding characters (β); B) A character may be fully (γ) or partially preserved (δ) yet not recognizable as part of a specific word (having no lexical value) because its context is lost: C) This character can be identified by other means, such as copies, similar texts, citations, etc. (ε). In other cases, the next criterion applies.
3. An invisible character can be restored with varying degrees of certainty in a lacuna by using internal criteria, such as discursive coherence or the existence of similar

28 The detailed data of the formula to obtain the Callimachus number in a document marked in TEI can be seen at https://glg.csic.es/Callimachus/Callimachus_formula.html.

texts that are not copies, citations, or paraphrases. In this context, a convention is established where five is considered the number of missing characters to assign greater or lesser plausibility to the reconstruction: a text reconstructed within a lacuna of five or fewer characters (ζ) receives a higher CN than one reconstructed in a lacuna of more than five characters (η). If the editor deems the reconstruction of the text in a lacuna hopeless or overly speculative, the next criterion is applied.

4. The algorithm then addresses illegible letters, writing traces, or lacunae. In each situation (i-viii), individual letter remnants receive a slightly higher CN than "traces," and these higher than a completely disappeared character: i) The lost character can be identified at least as part of a word (θ). ii) The lost character cannot be identified as belonging to a word (ι). iii) Mere writing traces exist where individual letters are not recognizable (κ). iv) The editor can determine the number of lost characters in a small space (λ). v) The space occupied by the traces (μ) in a line allows for establishing a not exact but approximate number of characters in the original. vi) The text has completely disappeared in a line, but the space allows for determining the precise number of characters in the original (ν). vii) The extent of the lacuna permits approximating the number of characters in the original text (ξ). viii) Traces of text across multiple lines exist where it is impossible to determine the number of lost lines (ο).
5. Finally, in cases where no writing traces are present and the lacuna occupies an extent that can only be estimated by indirect means, such as the probable extent of the support or proximity to the previous or subsequent text for discursive reasons, two situations are distinguished: i) The lacuna is indicated immediately before or after a recognizable text (π). ii) The lacuna is marked at the beginning or end of an unknown-length text (ρ). In both cases, the number of lost characters is tentatively estimated by assigning each lost line a number of characters based on the average number of characters per line in the same document[29] and estimating the number of lost lines according to the editor's indications or, in their absence, by other means.

In Table 1, a summary of the CN estimation, and the labels and attributes used in EpiDoc to indicate the type of situations referenced, can be seen. Using these labels, a computer algorithm can immediately estimate the CN of any text or fragment. The CN is useful for describing the state of a document in a way that allows for comparison with others, but also in other situations. For example, it can be a highly useful tool for creating

29 When using texts marked with EpiDoc to calculate the average number of characters (usually letters) per line in a document or fragment, Callimachus counts only the lines that have been fully preserved if there are more than four. Otherwise, it includes in the calculation of characters per line those lines where the editor has indicated the precise number of lost characters. When there are no fully preserved lines, and the editor has not been able to estimate the number of lost characters in a line, conventionally 7 characters are added to the number of letters in the longest preserved line.

corpora of fragmentarily preserved texts, as it allows for the selection of texts preserved in a similar state of preservation or readability. It can also be useful for creating treebanks: each sentence, and in fact each word or phrase, has its CN, and this can help select the texts to be analyzed or to assess the degree of significance of a particular analysis.

There are currently limitations and difficulties in applying the CN. The first is that it can only be applied to documents that contain some text. Otherwise, if, for example, the formula were applied to a text like the famous Egyptian papyrus from 2900 BC found in Saqqara in the tomb of Hemaka, which was never written on, we would obtain 0/0=indeterminate. More significant is the difficulty posed by the concept of 'document' when referring to incomplete texts, which one editor may understand as limited to the preserved text, and another may understand as referring to the original extent, resulting in a lower CN in the latter case for the same material support. To avoid these situations, a CN variant can be used that values only the text between the first and last fully preserved words (called Centered CN).

Finally, practice may recommend using a CN variation that involves squaring the CN of each letter and taking the square root of the result. This amplifies the distances between the best and worst-preserved documents.

Table 1: Criteria for determining the Callimachus Number (CN) corresponding to the readability (RCN) and preservation (PCN) of each letter or character in a text, the elements and attributes used by EpiDoc to mark each situation, and the value of each character in each of the situations. In the first column, the designation of the type of situation used in the text is shown. In situations such as (θ, ι, κ), the algorithm processing an EpiDoc document must take the context into account.

Type	State of the character	EpiDoc marking & context	RCN Value	PCN Value
(α)	visible character, part of a word	none	1	
(β)	unclear character, part of a word	`<unclear>`	0.9	0.7
(γ)	visible character, not part of a word	none	0.8	0.9
(δ)	unclear character, not part of a word	`<unclear>`	0.7	
(ε)	character supplied thanks to a parallel	`<supplied> @evidence="parallel"`	0.65	0
(ζ)	character supplied letter in a gap of less than 5 letters	`<supplied> @reason="lost", "undefined"`	0.6	0.1
(η)	supplied character in a gap of more than 5 letters	`<supplied> @reason="lost", "undefined"`	0.5	0.1
(θ)	illegible character, part of a word	`"illegible"`	0.4	0.3
(ι)	illegible character, not part of a word	`"illegible"`	0.3	
(κ)	«vestiges» of writing	`"vestiges"`	0.25	0.2
(λ)	illegible text; the number of missing characters can be counted	`<gap> @reason="illegible" @unit="letter" @quantity`	0.2	

(μ)	illegible text; the number of missing characters can be approximated	`<gap> @reason="illegible" @atLeast/ atMost`	0.18	0.2
(ν)	lacuna; the number of missing characters can be counted	`<gap> @reason="lost" @unit="letter" @quantity`	0.16	0.15
(ξ)	lacuna; the number of missing character can be approximated	`<gap> @reason="lost" @unit="letter" @atLeast/ atMost`	0.14	0.15
(ο)	illegible; number of characters unknown	`<gap> @reason="illegible" @unit="line" @extent="unknown"`	0.12	0.2
(π)	lacuna; near a word; number of characters unknown	`<gap> @reason "lost" @unit="line", "character" @extent="unknown"`	0.1	
(ρ)	lacuna; number of characters unknown	`<gap> @reason="lost" @unit="line", "character" @extent="unknown"`	0	

Bibliography

Bodard, G. (2010), *EpiDoc: Epigraphic Documents in XML for Publication and Interchange*, in *Latin on Stone: Epigraphic Research and Electronic Archives*, ed. by F. Feraudi-Gruénais, Lanham, 101–18.

Reggiani, N. (2019), *La papirologia digitale. Prospettiva storico-critica e sviluppi metodologici*, Parma.

Riaño Rufilanchas, D. (forthcoming), *Argos: un buscador en treebanks grecolatinos*, to be published in *Actas del XVI Congreso de la SEEC*, ed. by J. de la Villa *et al*.

Stolk, J. V. (2018), *Encoding Linguistic Variation in Greek Documentary Papyri. The Past, Present and Future of Editorial Regularization*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 119–38.

Vannini, L. (2010), *Review of Papyri.info*, RIDE – A Review Journal for Digital Editions and Resources 9, https://ride.i-d-e.de/issues/issue-9/papyri-info.

Klaas Bentein

Socio-semiotic, Multimodal Annotation of Documentary Sources

Digital Infrastructure in the Everyday Writing Project

> When all is said and done, we shall find that the activity of writing, like the activity of speaking, is a supremely social act. Simultaneously, I believe, we shall find that it is far more complex – and therefore more intriguing – than we have suspected heretofore.[1]

1 Writing: a complex, supremely social act

In a short 1989 contribution entitled *The Ethnography of Writing*, the cultural and linguistic anthropologist Keith H. Basso lamented the then current stagnation in the study of writing systems.[2] To reignite interest, Basso repositioned the subject within the framework of the ethnography of communication, viewing writing as a form of 'communicative activity' and directing attention to "the social and cultural factors that influence the ways written codes are actually used",[3] instead of focusing exclusively on the internal structure of these written codes.[4] Basso concludes his piece by emphasizing the supremely social, yet complex – and thus intriguing! – nature of writing. This emphasis is also mirrored here, under the headings of 'social semiotics' and 'multimodality,' forming a double helix that weaves through the remainder of the discussion.

Already in the 1960s, the study of linguistics had broadened from a cognitive, 'intra-organism' perspective to a more socially oriented 'inter-organism' perspective,[5] though to the exclusion of written sources, as lamented by Basso. From this time onwards, William Labov and other pioneers in the burgeoning field of sociolinguistics made significant advances in uncovering the intimate relationship between linguistic variation and the

1 Basso 1989, 432.

2 My work was undertaken in the context of the ERC Starting Grant project EVWRIT ("Everyday writing in Graeco-Roman and Late Antique Egypt. A socio-semiotic study of communicative variation", www.evwrit.ugent.be), a project which has received funding from the European Research Council under the Horizon 2020 research and innovation programme (Grant Agreement No. 756487). All hyperlinks last accessed on 21.7.2024.

3 Basso 1989, 426.

4 A more recent publication that seeks to redress the neglect of writing in the field of sociolinguistics is Lillis 2013.

5 The terminology is Michael Halliday's (Halliday 2010). For further discussion, see Bentein 2019b.

https://doi.org/10.1515/9783111070162-014

expression of social meaning, focusing specifically on spoken language. Starting from the early 1980s, however, scholarship increasingly came to recognize that the same forces that are at work in spoken conversation can be observed in written texts, too, and that in written sources, too, linguistic variation is a key enabling factor for what Jurgen Spitzmüller refers to as *social visibility*.[6]

That there are important connections to be made between linguistic features and the social context of writing has been amply discussed for antiquity, too, documentary sources such as papyri providing a privileged source for historical sociolinguistic analysis, given that they are transmitted directly from antiquity, have been preserved for an extensive period of time, often can be dated, and are contextually diverse, ranging from scrap papers and shopping lists to official petitions and imperial edicts.[7] At the same time, papyrologists have stressed other, non-linguistic aspects of documentary sources that transmit social meaning; such extra-linguistic aspects of writing convey information that goes beyond the literal meaning of the text, elucidating elements such as the relationship between the author and their intended audience, as well as revealing their cultural affiliations. Petra Sijpesteijn, for example, has noted in her monograph on early Arabic papyrus letters that elements such as handwriting, linguistic register, and writing material all transmit indirect social messages concerning hierarchy, authority, and power relations.[8] Most forcibly and programmatically, Jean-Luc Fournet has argued for the recognition and establishment of what he calls "paléographie signifiante", noting that "l'analyse matérielle d'un document peut être porteuse de sens",[9] not only when it comes to text type, but also with regard to the socio-cultural context of writing, and the provenance of the document.

Scholars working in the field of 'social semiotics', a discipline that attempts "to describe and understand how people produce and communicate meaning in specific social settings",[10] first developed the concept of *multi-modality* to describe the different semiotic 'modes' or 'resources' that are used to make meaning besides language, such as pictorial, gestural, musical, choreographic, and most generally actional resources.[11] Remarkably, however, this discipline remained restricted to the analysis of modern-day texts, social semioticians showing little interest in documents from the past – thus resembling sociolinguistics in its initial stages. From 2018 to 2024, a large-scale European-funded project was conducted at Ghent University, entitled "Everyday writing in Graeco-Roman and Late Antique Egypt. A socio-semiotic study of communicative variation"

6 Spitzmüller 2013, 1.

7 See e.g. Logozzo 2015 for formulaic expressions in the Zenon archive, Bentein 2017 for complementation patterns in the Roman and Late Antique period, to name but some studies.

8 Sijpesteijn 2013, 255.

9 Fournet 2007, 353.

10 Kress – Van Leeuwen 1996, 266.

11 Lemke 1998.

(EVWRIT),[12] the main purpose of which was to study the communicative choices made by writers in their papyrus documents, and how these communicative choices can be related to the broader context of communication, thus incentivizing the establishment of a new 'historical socio-semiotic' approach to communication practices in antiquity more broadly.[13]

In this chapter, I outline the digital infrastructure that was developed in the context of the project to capture both the 'supremely social' and 'complex' nature of writing, to borrow Keith Basso's description. The chapter is structured as follows: I will start by outlining how we designed our database, accommodating the needs of the researchers in the Everyday Writing research team as well as taking into account the latest findings in communication studies (§2); I will then go on to explain how we operationalized each of the annotation layers in the database (§3), and which tools we developed to query the extensive set of annotations that we created (§4).[14] Before making some concluding observations about short- and long- term plans and possibilities (§6), I briefly illustrate the digital infrastructure that we developed through two test cases, thereby distinguishing between distinct, but related branches of research (§5).

2 A platform for socio-semiotic, multimodal annotation

Initial socio-semiotic studies by pioneers such as Gunther Kress and Theo van Leeuwen heavily relied upon the foundational linguistic work by the late M.A.K. Halliday, who had characterized language as a 'social semiotic', recognizing already at a very early stage that

> There are many other modes of meaning, in any culture, which are outside the realm of language. These will include both art forms such as painting, sculpture, music, the dance, and so forth, and other modes of cultural behaviour that are not classified under the heading of forms of art, such as modes of exchange, modes of dress, structures of the family, and so forth. These are all bearers of meaning in the culture. Indeed we can define a culture as a set of semiotic systems, as a set of systems of meaning, all of which interrelate.[15]

One of Halliday's key insights was that communication is not only *multimodal*, but also *polyfunctional*, whereby he hypothesized the existence of three distinct types of 'meaning', which he referred to as 'ideational' (construing our experience of the world and our consciousness, e.g. 'apple' = fruit for eating), 'textual' (organizing discourse and creating

12 See further www.evwrit.ugent.be.
13 For which, see now Bentein – Amory 2023.
14 Neither of these tools is publicly available at the moment, but they should be launched in the foreseeable future.
15 Halliday – Hasan 1989, 4.

continuity and flow in texts, e.g. “it’s raining, therefore I will take an umbrella”, *therefore* indicating a consequential relationship between two clauses), and ‘interpersonal’ (enacting personal and social relations, e.g. “you could try this”, *could* indicating a suggestion or possibility). Early socio-semiotic work used these three areas of meaning-making to explore other semiotic modes, visual communication in particular.

While greatly advancing our insights into the multi-modal nature of communication practices, socio-semiotic approaches towards multimodality have been criticized in various regards, among others because of the fluidity of key concepts such as ‘semiotic mode’ and the lack of a clear analytical framework. One of the central issues in multimodal research since then has been the development of a model that can provide a framework for the description of multimodal documents, and that is capable of disentangling the various modes that play a part in making meaning, and how they come together. One such model is the so-called *GeM* (‘Genre and Multimodality’) model, which was developed by John Bateman and his associates over the course of the first two decades of the twenty-first century.[16] Figure 1 shows a schematic representation of the GeM model:

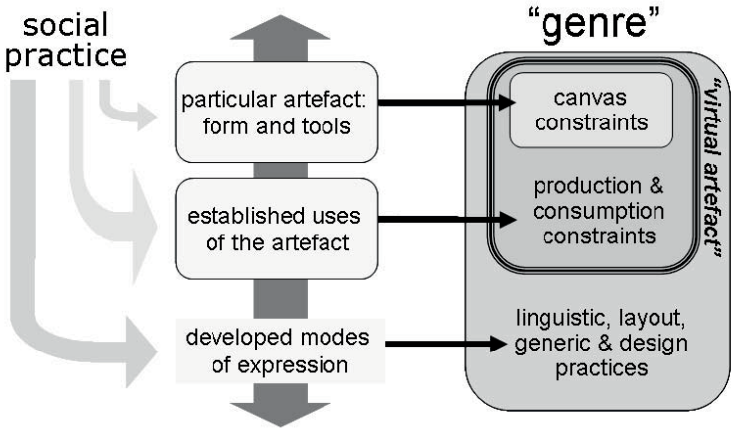


Fig. 1: Bateman’s *GeM* model (from Bateman 2008, 16).

As can be seen, this approach positions the creation of multimodal artifacts within a network of social practices. From the perspective of production, it views documents as being generated and utilized amid a set of restrictive forces, which are categorized into three distinct types. Documents created are subject to limitations not just from the material or surface employed (*canvas constraints*) but also from the technologies utilized in their creation, like limitations on color availability (*production constraints*), as well as

16 Relevant publications include Bateman 2008; Hiippala 2016; Bateman – Wildfeuer – Hiippala 2017.

the purposes they are meant to serve (*consumption constraints*). These three types of constraints give rise to what is called a 'virtual' artefact, to which certain developed modes of expression are then applied.

These developed modes of expression are analyzed and described by making a strict distinction by various types of structure or 'layers',[17] including a 'base layer, 'layout layer', 'rhetorical layer', and 'navigation layer'. Space does not allow to expand on these different layers, but let me note here the importance of distinguishing between visual and rhetorical structure. This distinction may seem trivial, but it is not: it can throw light, for example, on what constitutes the 'opening' of Greek documentary sources, about which there does not seem to be a consensus. According to one point of view, the body of Greek documentary letters starts with a health wish and *proskynema* formula, whereas the opening (prescript) consists of the name of the initiator and addressee in combination with a greeting formula.[18] This view, however, does not explicitly distinguish between visual and rhetorical structure: visually speaking it may be true that the name of the initiator and addressee together with a greeting verb is set apart, but that does not mean that rhetorically speaking the health wish and *proskynema* formula do not belong to the opening, and in fact it is quite common for interpersonal formulae to cluster together at the beginning and end of communicative acts.[19]

The database that we created in the context of the Everyday Writing project does not blindly follow Bateman's GeM model, but takes inspiration from it, especially in the distinction of various layers of description. Due to the nature of the sources we are working with, the earlier history of our database, and the project's objectives, our own database has some specific points of emphasis, such as, among others, socio-pragmatic annotation and text segmentation. Structurally, the Everyday Writing-database consists of five main annotation areas, which we refer to as (i) *metadata*, (ii) *materiality*, (iii) *text structure*, (iv) *base annotations*, and (v) *languages*. I will further outline how we operationalized each of these annotation areas in the next section (§3), but let me point out for now that they involve different *units* of analysis: metadata and materiality relate to documents in their entirety, text structure to larger-scale segments of these documents, and base annotations and languages to more specific segments.

In terms of database history, a first version of the database, shown in Figure 2, was developed in Microsoft Access in the academic year 2013-2014,[20] in the context of a post-doc-

17 See Bateman 2008, 19 for an overview.

18 Luiselli 2008, 692, 700.

19 Compare e.g. Bernhart – Wolf 2006 on 'framing borders'.

20 The developer was my father, Gilbert Bentein, who also created another MS Access database that has known a similar trajectory, the *Database of Byzantine Book Epigrams* (Ricceri – Bentein – Bernard *et al.* 2023), which is online available at https://www.dbbe.ugent.be. I want to express my profound gratitude to my father for the substantial time and energy he invested in creating these databases, which have been pivotal to my academic career.

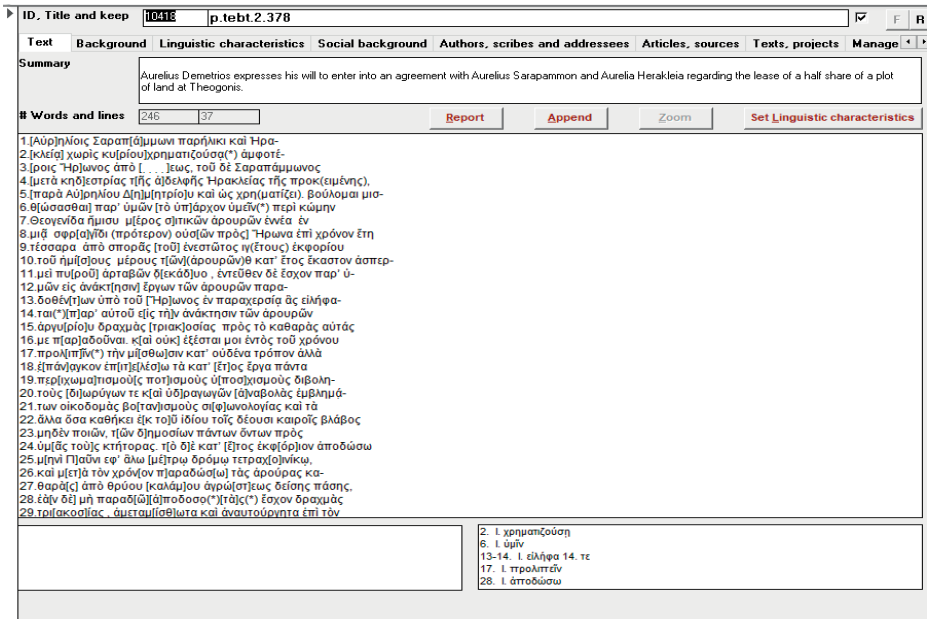


Fig. 2: Everyday Writing database (2013-2014 version).

toral research project on linguistic variation in documentary sources funded by the Flemish Fund for Scientific Research and the Belgian American Educational Foundation.[21] In a next stage, the Access database was expanded and further enriched with metadata with the help of KULeuven's Trismegistos team,[22] and converted to FileMaker, which allows it to be used by multiple users simultaneously. Throughout, emphasis was put on 'interdirectional' texts, in other words texts with a clear initiator-receiver structure, such as letters, petitions and contracts,[23] the idea being that such texts have more inherent motivation for communicative variation, that is, they create more opportunity for intersubjective positioning than for example lists and accounts do. The Everyday Writing project specifically focuses on a subset of the material, namely documents dating to the Roman and Late Antique period (I-VIII AD) that either originate from or were found in Middle Egypt.

21 Earlier publications such as Bentein 2015a, 2015b, 2017 rely on this version of the database.

22 I want to thank Mark Depauw and Tom Gheldof for their invaluable help and support.

23 Not all contracts are interdirectional strictly speaking; nevertheless, they have all been included.

3 Operationalizing the annotation layers

In what follows, I will briefly outline how we operationalized the different annotation layers in the Everyday Writing database, that is, which specific features we decided to annotate. Let me stress from the beginning the functional-paradigmatic orientation of the database:[24] as we do not begin from specific formal values, but structure the database in annotation areas, which are further divided in language-independent variables with systems of choice which are themselves language-independent to various extents,[25] the database can in principle be used by scholars working on any (ancient) language or corpus. In fact, the database has been used for a doctoral project about the language of Latin inscriptions,[26] as well as for a post-doctoral project about the language of early Arabic papyri,[27] and another post-doctoral project about the linguistic features of early Post-classical inscriptions.[28] This in turn makes it possible to engage in cross-corpus and cross-cultural comparison of textualization practices, a field of research that we have engaged in only to a limited extent.[29]

3.1 Texts

Texts is the only annotation area that is shared by all team members, and that is used for entering both essential textual data, as well as for annotating documents for their socio-pragmatic characteristics. With essential textual data, I mean:

- the texts themselves (whether Greek, Latin, Coptic or Arabic), which, for papyri at least, to a large extent derive from the *Duke Databank of Documentary Papyri* (DDbDP);
- basic metadata, again copied from the DDbDP and enriched with information that was shared by Trismegistos, such as *keywords*, *archive*, *place and time of writing*, and *find place*;
- basic material and linguistic data, such as translation, material substrate, language, script, and production stage;

24 One can compare this to the paradigmatic organization of the systemic-functional language model proposed by Halliday, which recognizes different levels or *strata* that stand in a realizational relationship to each other (see further Halliday and Matthiessen 2013).

25 E.g. Materiality > Orientation > *horizontal* vs. *vertical* or Language > Syntax > ComplementationContext > *preposed* vs. *postposed*.

26 See https://research.flw.ugent.be/en/projects/contribution-inscriptional-evidence-analysis-vulgar-latin-vowel-system-ranging-republican.

27 See https://research.flw.ugent.be/en/projects/chaos-order-quantitative-approach-variation-arabic-papyri-7th-9th-centuries-ad.

28 See https://research.flw.ugent.be/en/projects/sociolinguistic-variation-ancient-greek-dialects-mapping-contact-between-doric-and-koine.

29 See the pilot study by Bentein – Kootstra forthcoming.

– project-related data, such as the unique EVWRIT ID, and one or more scholarly projects each document forms part of, such as the Everyday Writing Project.[30]

While the Trismegistos platform nowadays contains a 'content' field with an indication of a document's text type, at the time when we started the project, no such information was available.[31] We therefore decided to set up our own text typology, which recognizes text types at three hierarchical levels, called *hypertype* (e.g. LAW), *type* (e.g. 'contract') and *subtype* (e.g. 'contract of lease').[32] We also included a field where to annotate the original text label mentioned in the text, if that is available, again with the option to indicate subcategories (e.g. ὁμολογία "agreement, contract" > διαλυτικὴ ὁμολογία "agreement of settlement").

Besides providing essential textual data, we also use the 'texts' area for socio-pragmatic annotation, that is, for adding information about the nature of the communicative act. This sort of annotation, which is rather time-consuming and which cannot be derived from any other platform, naturally focuses on the participants to the communicative act.[33] For each communicative participant, we enter information such as his/her *name*, *patronymic*, *alias*, *gender*, *age*, *education (literacy)*, *occupation*, *social rank*, *domicile*, *honorific epithet*, and *communicative role*.[34] Each person is attributed a unique ID, which we use in conjunction with the Trismegistos Person ID.

For each person mentioned in the text, the goal is to add as much social information as possible, including not only *absolute* social information such as people's gender or domicile, but also *relational* social information,[35] which concerns people's communicative intentions and relations in specific communicative acts. To this end, we have added fields in the database that specify the *communicative goal* (the goal an initiator wants to achieve with his/her communicative act, e.g. giving an order, making a request, thanking someone, etc.); the *social distance* between the initiator and receiver (the degree of formality of the interaction); and *the agentive role* (the extent to which the relationship is hierarchical or not, e.g. subordinate to superordinate).[36] For each of these three fields, we have found it useful to work with categories and subcategories, so as to maximally structure the infor-

30 One document can be assigned to multiple projects.

31 The DDbDP has a 'subjects' field with keywords for each text. While this field often contains useful information with regard to a document's text type, the information is not presented in a systematic way.

32 A similar approach is pursued by the Grammateus project (https://grammateus.unige.ch), which recognizes *types* on the basis of general communicative goals (e.g. 'transmission of information'), *subtypes* (e.g. 'declaration') and *variations* (e.g. 'census declaration').

33 Unlike Trismegistos People, we do not attempt to cover *all* persons mentioned in a text.

34 Three roles are central, namely those of the 'initiator', 'receiver', and 'scribe'. At the same time, we also recognize other roles such as those of 'witness', 'signatory', 'intermediary', 'consenter', 'legal representative', 'copyist', etc. *Saluters* and *salutees*, which play a prominent role in private letter writing, are, at the moment, not included.

35 For these two main types of social identity information, see Bentein 2019, 145.

36 For more elaborate discussion of social distance and agentive role, see Bentein 2017, 22–8.

mation. For agentive role, for example, we recognize four main categories ('equal to equal', 'family member to family member', 'subordinate to superordinate', and 'superordinate to subordinate', each of which can then be further subdivided, e.g. for family members > 'daughter to mother', 'husband to wife', 'uncle to nephew', etc.).

3.2 Materiality

Our materiality area[37] is similar to 'texts' in the sense that it involves information at the level of the entire document. This area of the database contains two types of materiality information, related to the disposition of the material substrate used for writing on the one hand and the materiality of the actual writing on the material substrate on the other. Annotation fields for the former include among others *writing side* (e.g. 'recto'), *orientation* (e.g. 'horizontal'), *form* (e.g. 'roll', 'sheet') and *writing direction* (e.g. 'perfibral'). Annotation fields for the latter include *number of lines*, *letters per line*, and *columns*.

A third type of information that we annotated relates to measurements of the material features of written documents, such as *height*, *width*, *margin size*, *line height*, *interlinear space height*, and *kollemata size*. Unfortunately, these measurements, which are key to the interpretation of the materiality of a document, are often not included in papyrological editions, especially older ones. We have therefore developed a new tool, called the 'measurement tool', which is able to perform measurements on the basis of digital images. As this tool is further discussed in Serena Causo's contribution to this volume, I will not go much further into it here. Suffice it to say that the tool is able to capture the required measurements on the basis of a number of manual manipulations that the scholar has to make on analytical units of various size, that is, the entire document, the actual text (the 'positive' space), a representative line, and a representative interlinear space. Calculating these different types of measurements is done through bounding box annotation, as shown in Figure 3, and is based on a unit of scale that typically derives from the ruler that is included in the digital image.

Of course, the procedure that we developed to retrieve these measurements is quite time-consuming, and, certainly for smaller-scale measurements, limited, as it is based on a 'representative' unit.[38] Moreover, when drawing boxes the scholar is faced with a number of difficulties as to what part of the text should or should not be included, which of course influences the measurements that we obtain. That being said, the measurement tool fills a substantial gap in current scholarly knowledge.

37 The materiality area was developed in collaboration with Serena Causo and Febe Schollaert.

38 For example, interlinear space is based on a single interlinear space that we measure in the document and that we consider representative.

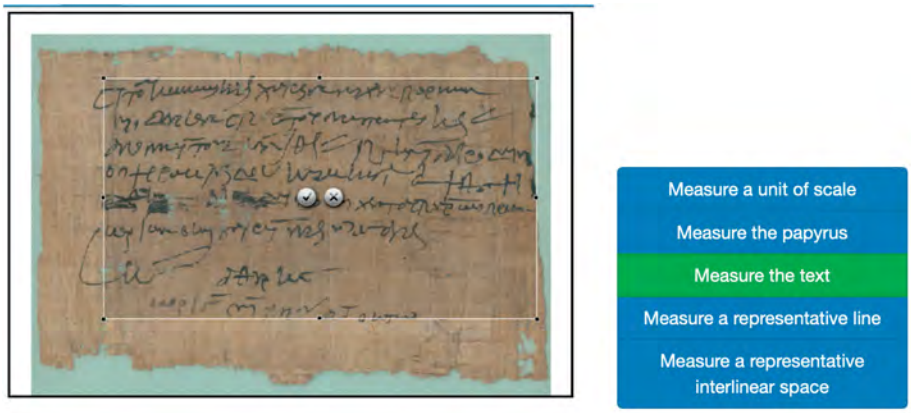


Fig. 3: Measuring the size of the text with the measurement tool.

3.3 Text structure

Unlike our two previous database areas, text structure goes below the level of the document in its entirety, looking into its internal organization, both visually and linguistically: we do so by recognizing four types of text structure, which we refer to as *generic structure* (e.g. is there an opening, body, and closing?), *lay-out structure* (e.g. is the opening visually set apart?), *handwriting* (e.g. are there multiple hands at work?), and *levels* (e.g. is one text embedded in another?).[39]

Generic and lay-out structure have the same type of general design:[40] the assumption is that both are structured hierarchically, and that this hierarchical structure can be described by making use of the same or at least similar types of segments, which we call, from large to small, *part*, *unit*, *element* and *modifier*.[41] Parts and units are the largest linguistic and visual constituent elements of a document. Each text consists of at least one visual and linguistic part, but usually out of multiple such parts. In fact, many texts have a threefold structure, which is called 'opening', 'body', 'closing' for generic structure, and 'initial part', 'main part', and 'final part' for layout structure.[42] One part may consist out of multiple units: for example, the main part may be divided into two units

39 Generic structure was developed in collaboration with Marta Capano and Fokelien Kootstra; lay-out structure in collaboration with Serena Causo and Fokelien Kootstra; handwriting in collaboration with Yasmine Amory; and levels in collaboration with Serena Causo and Gianluca Bonagura.

40 For more elaborate discussion of lay-out structure, see Bentein – Kootstra forthcoming.

41 One should not consider these four segments exhaustive. The line, too, may be thought of as a central local unit, with a visual function that is comparable to that of the clause (complex) rhetorically speaking (Crystal 1979).

42 So as to avoid confusion, we adopt different terminology for generic and visual annotations at the highest level of segmentation.

by a *paragraphos*. Whereas parts and units as segments serve global reading strategies, elements and modifiers are relevant at a more local level of reading. While there may not be an exact correspondence between generic and lay-out structure, we have found that working with the same types of segments opens the door to explicit comparison of different types of textual structure.[43] Segmenting the document is an important part of the generic and lay-out structure annotation process, but the more time-consuming part of these annotations is indicating which linguistic and visual cues justify the proposed segmentation. For lay-out structure, for example, we recognize as many as eight different systems of visual cueing, which, for any given segment, can be at work simultaneously. For generic structure, on the other hand, formulae play a key role, together with other linguistic features.

Besides the generic and lay-out structure of the document, the Everyday Writing database allows to enter paleographical information about the handwriting(s) found in the text. For this purpose, we have developed a set of criteria that can be used to describe handwritings, including such fields as *script type*, *degree of formality*, *expansion*, *slope*, etc., basing ourselves on earlier work by Theo van Leeuwen, who argued for the need to include typography in the broader field of multimodality, and proposed a system of distinctive typographical features which can be used for the typographical analysis of letterforms.[44] For each hand, we also indicate how it is related to typographical features such as *punctuation*, *accentuation*, *word splitting*, *abbreviations*, and *corrections*, limiting ourselves to a simple 'yes' or 'no' annotation.[45] An example of an annotated hand is shown in Figure 4.

Information about the number hands in a text is taken from the DDbDP, but sometimes it is necessary to reduce or expand the number of hands, especially for texts that were edited a long time ago, when hand shifts were less systematically reported.[46] We have therefore made it technically possible to adjust hand shifts with respect to the way that they are indicated on the DDbDP. Each hand can also be linked to the persons that play a role in the communicative design of the text and that were annotated in the database under 'texts'. Even though it is more often than not unclear who wrote (sections of) the text, the linking of paleographical information and metadata should make it possible, in time, to make broader generalizations, for example about the relation between social class and script type.

43 One divergence we had to introduce between generic and lay-out structure concerns the introduction of subtypes of units (called 'subunits') and of modifiers (called 'complex modifiers'). That generic structure should have a more complex hierarchical organization is in itself not a surprise, given the complexity of language as a semiotic system.

44 van Leeuwen 2006. See further Amory forthcoming for our application to documentary sources.

45 Further details can then be added in the typographical section, see §3.4.

46 See e.g. Sarri 2016 on the need to correctly identify handshifts in letters.

Fig. 4: Handwriting annotation for Eudaimonis.

One factor that considerably adds to the complexity of annotating documentary sources, socio-pragmatically and otherwise, is that many of them – formal documents in particular – tend to have a complex structure, something which is abstracted away from in current papyrological portals, with e.g. a single text type being assigned to a textually complex document. In order to deal with this textual complexity, we created the option to recognize multiple 'levels': a document with one main text and another text in attachment can be said to have two 'levels', each of which receives its own socio-pragmatic annotation in the database. Many documents are not limited to two such levels: in our database, there are documents with more than ten levels. Moreover, relations between texts on a single papyrus need not be hierarchical: one can think, for example, of documents containing multiple private letters that are not embedded in each other, but rather placed together for practical reasons (e.g. letters from distinct persons for a single addressee).[47] Not in all cases is such a textual relationship equally clear or explicit (e.g. in the case of registers containing unrelated texts), which raises the question of what defines a single text/document, a topic I will not delve into further here.[48]

3.4 Base annotations

Base annotations are made at the lowest level, usually involving words or combinations thereof. These annotations are created by selecting any given part of the text, and filling

47 In the database, we have created a typology of levels-relations.

48 The Trismegistos platform provides some useful guidelines at https://www.trismegistos.org/about_how_to_cite.php.

out further information in one or more database fields. Each such annotation receives its own unique annotation ID; its position in the text is established in multiple, complementary ways: for each annotation, the first and last line on which the annotation is found are recorded, as well as the first and the last character of the annotation, and finally also the TM Words ID of the first and last word of the annotation. This positioning process allows us to highlight the relevant annotation in the text, both in the database and on the website, as I will discuss further in §4.1.

In the database, we recognize two broad types of base annotations, which we refer to as linguistic and typographical respectively,[49] further splitting up linguistic annotations in *morpho-syntactic*, *orthographic*, and *lexical* annotations. Let me start the discussion of typographical annotations[50] here with a brief note about the term 'typography', which some scholars may consider anachronistic. Whereas this term is sometimes associated with printed text, it is now increasingly being used to refer to the visual organization of written language however it is produced.[51] Contrary to linguistics, typography does not have a formal and established descriptive tradition,[52] which means that we had to decide ourselves which descriptive aspects to include. After some deliberation, we decided to focus on two aspects: first, we included features of textual presentation that were made by the scribe either during or after the writing process, and which can be grouped under the heading of 'text management': these include *word splitting*; *abbreviations*; and *deletions*, *insertions*, and *corrections*.[53] Whereas for handwriting we described these same fields in terms of a binary (yes/no) distinction,[54] under 'typography' each specific abbreviation, word split, etc. can be annotated, and further information can be added. A second group of typographic annotations relates to features that aid the reader in the interpretation of individual words (diacritics) or larger constituent parts of the texts (lectional signs). This includes such features as *accentuation*, *punctuation*,[55] *symbols*, and *vacat*.

Evidently, both categories, text management and diacritics/lectional signs, are broad categories that require an enormous amount of annotation work: as the Everyday Writing project did not have the (wo)manpower to engage in this task, we have limited ourselves to automatically collecting all relevant information from the editions of the texts in the DDbDP, and manually going through the annotations that were collected to add information. This approach saves a significant amount of time, though it does have notable disadvantages, too, since, as already noted, the online editions – especially those

49 Typology was developed in collaboration with Yasmine Amory.

50 The database's typographical section was developed in collaboration with Yasmine Amory.

51 Walker 2001.

52 Walker 2001, 17, 23.

53 We consider corrections to be a combination of a deletion and an insertion.

54 See 3.3.

55 The distinction between 'accentuation' and 'punctuation' is less straightforward than it may seem, but I will not go further into this here.

taken from older works – often do not contain any or all relevant typographical information; moreover, the DDbDP may not contain an exact copy of the printed edition when it comes to such information.[56] Much of the typographical work therefore remains to be done in the future, in collaboration with other projects.

Large part of the linguistic annotations in the Everyday Writing database are dedicated to clause 'linking' or 'combining', that is, how clauses are related to each other in discourse.[57] This includes what in linguistic scholarship are called *co-ordination*, *complementation*, *relativization*, and *adverbial subordination*. Each annotation that is made about a selected part of the text has a threefold structure: *form* relates to the standardized form of the morpho-syntactic feature that has been annotated; *content* relates to the semantic value of the morpho-syntactic feature; and *context* to the sentential or broader textual context in which the feature can be found. Besides morpho-syntax, other linguistic levels can be annotated, too: this includes orthographic and lexical annotations, which have a parallel structure in the database. I will not go any further into these here, for reasons of space.

3.5 Languages and scripts

As I already noted in the introduction to this section, the database's functional-paradigmatic orientation allows it to be used by scholars working on any (ancient) language, enabling cross-corpus and cross-cultural comparison. The languages and scripts area in the database does not relate to the use of multiple languages and scripts *across* documents, but rather delves into such variation *inside* documents.[58] Whereas many texts are written in a single language/script,[59] one often can find switches in one and the same text between languages and/or scripts, ranging from individual letters to larger passages.

In order to automatically detect multilingual documents, we have developed a specific tool, called the 'character recognition tool', which processes each individual character of a transcribed document, and assigns it to a certain script. The tool calculates both the number of characters of each script in a document, as well as the percentage of characters in a certain script. In Figure 5 below, for example, the tool has detected 115 Latin characters, versus 368 Greek characters, and automatically calculated the relative weight (percentually) of the two languages in this document (76.2% Greek vs. 23.8% Latin). While this greatly facilitates searching for multiscriptal documents, the results are of course limited by the nature of the transcription: if the transcription reads 'x lines

56 As is e.g. the case with the Christian symbols, as also noted by Carlig 2020, 272.

57 See e.g. Buijs 2005 on clause combining in Classical Greek literature.

58 The languages area was developed in collaboration with Antonia Apostolakou.

59 It is important to keep apart 'language' and 'script': one language can be written in multiple scripts (e.g. Greek in Greek characters vs. Greek in Latin characters).

in language y', the tool will retrieve supposedly 'Latin' characters, and not the actual language that is used.

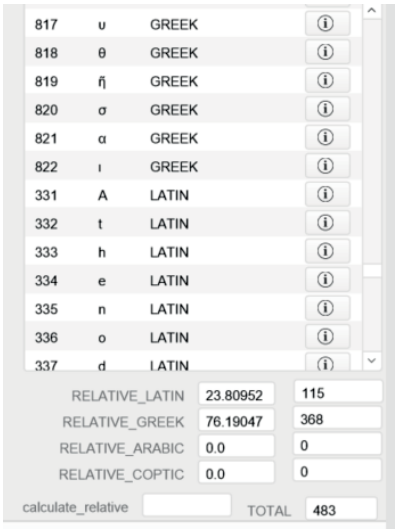


Fig. 5: Character recognition tool (ChLA XLI 1187 [298-300 AD] = TM 18367)

In the languages and scripts section, one can, besides giving a general indication of the languages and scripts used in a document, also select specific portions of each document that can be further annotated for categories such as *codeswitching* (every change from one language to another within the same document) and *transliteration* (a section of text in one language written in a 'non-standard' script). Each of these categories is further split up a in a number of fields, such as *type* (e.g. 'intersentential', 'intrasentential'), *rank* (e.g. 'noun phrase', 'verb phrase', etc.), *formulaicity* (e.g. 'formulaic', 'non-formulaic'), and *domain* (recurrent thematic elements, e.g. 'date', 'signature', 'personal name'). The selection and annotation method that we use is the same for base annotations.

4 Displaying and querying annotated information

Having discussed how we annotate information in the Everyday Writing project, I will now proceed to outline two new tools that we have developed to analyze and query the annotated information. These tools are the *Everyday Writing website* (§4.1) and the *Everyday Writing data exploration tool* (§4.2).

4.1 The Everyday Writing website

While the FileMaker database that we developed is ideally tailored for in-depth multi-user annotation work, it is less optimal for displaying and querying the entirety of the annotated information. For this reason, we created a project-dedicated website.[60] To this end, the non-SQL compliant FileMaker database was migrated to a modern relational PostgreSQL database system, which involved the creation of a new database model, as well as the parsing, cleaning and automated importing of all FileMaker data into the new database infrastructure. In addition, an advanced search service was developed utilizing Elasticsearch and the PHP Symfony framework to facilitate data aggregation and efficient data retrieval. Finally, a rich search and viewing application was constructed using VueJS. At this moment, information from the FileMaker database is updated to the PostgreSQL database system on a daily basis.

The general structure of the website follows that of the FileMaker database, as shown in Figure 6, where a search is made for epistolary communication between family members, though base annotations (§3.4) has been split up in two parts, orthography/typography, and lexicogrammar respectively. Each section of the website (*Texts*, *Materiality*, *Text Structure*, *Languages*, *Orthography and Typography*, *Lexicogrammar*) is similarly structured in that it contains on the left-hand side an extensive list of selection criteria (filters), and on the right-hand side a list of retrieved texts (or, for some website areas, a list of annotations made in those texts).

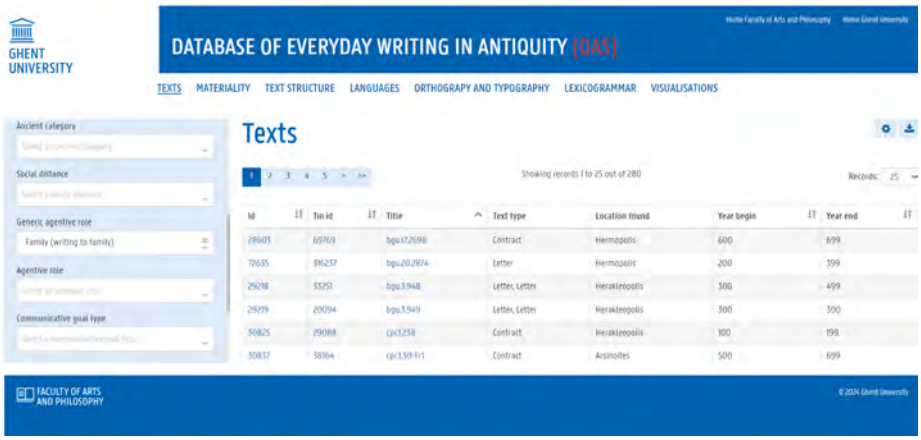


Fig. 6: Text search for generic agentive role 'family'.

60 The Everyday Writing website was developed in close collaboration with Frederic Lamsens (Ghent Centre for Digital Humanities).

What makes the Everyday Writing website a powerful tool for multimodal and socio-semiotic study is that it is incrementally structed, so that filters from all sections can be combined with each other: Materiality also contains the 'Texts' filters, Text Structure the 'Texts' and 'Materiality' filters, and so on. This allows for very complex searches of the type,[61]

> all instances of iota adscript (~ Orthography)
> in health wishes (~ Text Structure > Generic Structure)
> in visually distinct openings (~ Text Structure > Layout structure),
> in ostraca (~ Materiality)
> written by one family member to another (~ Texts)

Each time that an additional filter is added, the numerical values that are initially displayed (in our case for iota adscript) are adjusted accordingly.

While for Texts, Materiality and Text Structure, query results are returned in the form of lists of texts, for the other three areas – Languages, Orthography and Typography, and Lexicogrammar – actual annotations are displayed, besides the text, with an indication of their absolute and relative frequency ('instances in text' vs. 'frequency per line'), as shown in Figure 7, which displays results for 'asyndetic parataxis' under *ComplementationForm*. This allows the user to structure the analysis of the relevant texts, by either starting with the document with the lowest or highest frequency of the feature that is searched for. Once a text is clicked on, one does not need to return to the results list, but can use arrows to directly browse to the next relevant text.

Bringing together a multitude of complex information in a user-friendly way is a strength of the website at the level of individual texts, too. Inspired by Perseus' *Scaife Viewer*[62] and other such tools, we have created a layout with the text on the left-hand side, and an information pane on the right-hand side with basic information about the text in terms of metadata, imaging, materiality and people, and with several options to enrich the display of the text, for example allowing users to switch on or off the display of the translation or the critical apparatus. Another option that users have is to highlight one or more annotation types[63] with respect to the edited text, such as base annotations or language and scripts. Users can furthermore visualize aspects of text structure,[64] such as the generic structure, levels, layout structure and handwriting. Annotations and text structure features are interactive, which allows the user to quickly grasp what sort of information has been entered in the database.

61 There is an 'advanced mode' option that can be switched on for users who are interested in such complex searches. Other users can decide to display less filters.

62 Available at https://scaife.perseus.org.

63 For base annotations, see §3.4 above.

64 For text structure, see §3.3 above.

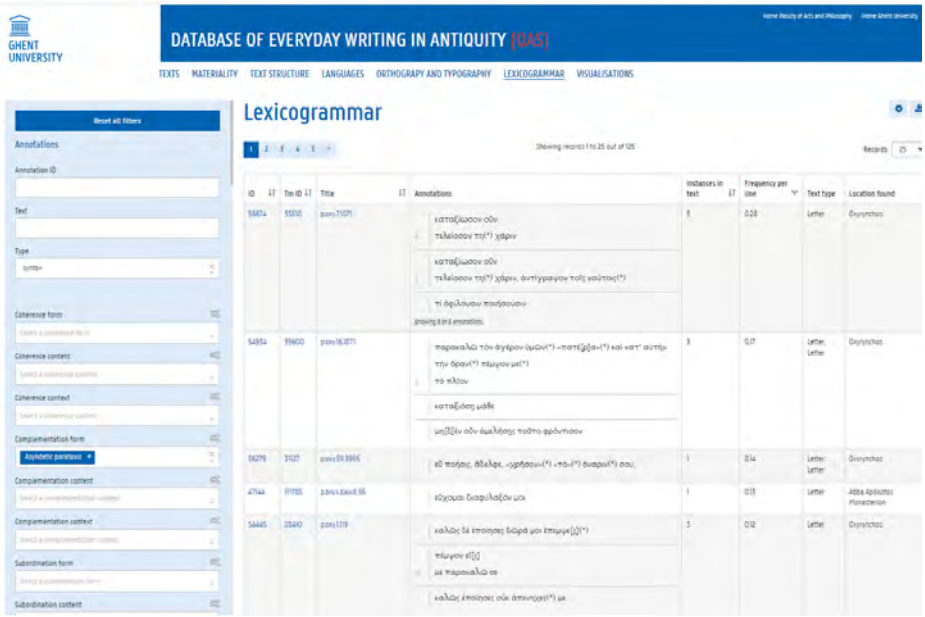


Fig. 7: Display of results, with indication of number of annotations and frequency per line.

Presenting such a vast array of information while ensuring clarity and maintaining an overview presented a significant challenge. Figure 8 shows some of the design and lay-out decisions that we have made: generic structure and layout structure are kept apart, with major segments structuring the visualization, while handwriting is displayed as a colored vertical line. Finer-grained generic and lay-out structure annotations are indicated through different types of underlinings, whereas colored boxes indicate different annotation types (typography annotations in red, language annotations in blue, syntax annotations in green etc.).[65]

4.2 The Everyday Writing data exploration tool

The website that we have created allows the user to engage in socio-semiotic, multimodal research, through extensive querying and visualization facilities. While greatly facilitating qualitative, and to some extent also quantitative analysis, it remains somewhat difficult to detect frequency patterns and correlations between annotation fields. In order to supply the user more directly with quantitative information, we decided

65 One can note that the colored boxes in the information sidebar indicate the numerical frequency of annotation types with respect to the text that is being displayed.

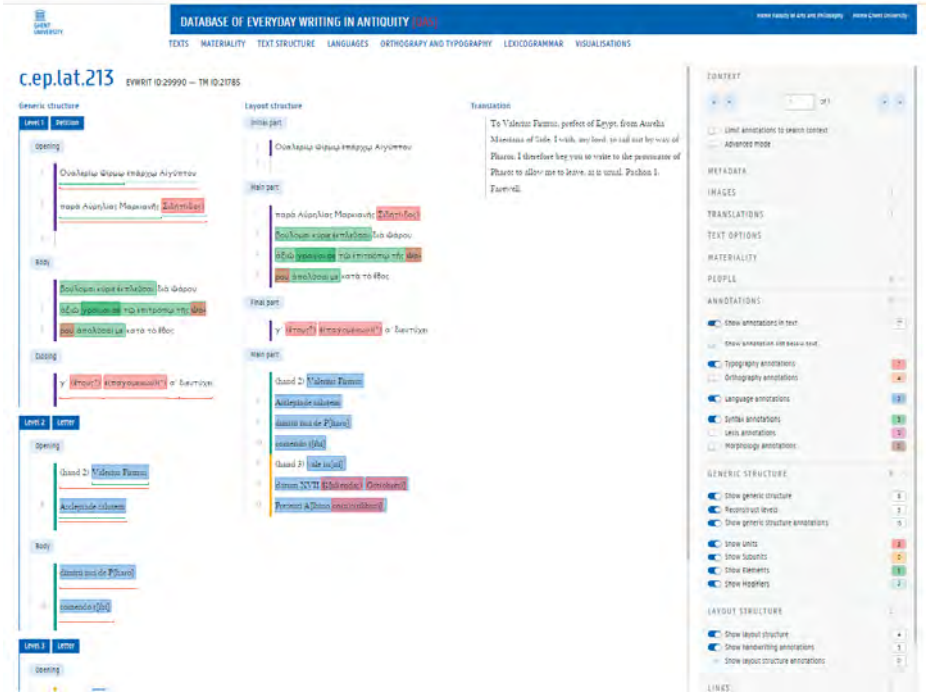


Fig. 8: Visualization options for individual texts

to create a 'data exploration tool' that is complementary to the Everyday Writing website.[66] This tool was created in R, a type of programming language that is commonly used in the humanities and elsewhere for the purposes of data analysis and visualization. We specifically leveraged the capabilities of the RShiny library, an R package that not only simplifies the creation of interactive web applications from R but also supports user input and interactive visualizations within these applications.

This data exploration tool is structured somewhat differently from the Everyday Writing website (and database) in the sense that organizationally it distinguishes between (only) two areas, which are called 'corpus overview' and 'feature overview' respectively, with the former focusing on metadata and materiality, and the latter on text structure, base annotations, and languages and scripts. Both areas allow the user to easily retrieve numerical data with respect to the variables included in the Everyday Writing database. Figure 9 displays a search for archives included in the Everyday Writing corpus, which are displayed in the form of a bar chart visualization, a graphical display that uses bars of varying lengths to represent data. For each such plot, one can addition-

66 The Everyday Writing data exploration tool was developed in close collaboration with Thomas Koentges (YouSayData).

ally generate a table which lists all relevant texts (and, in some case, annotations), with links to major papyrological platforms such as the DDBDP and Trismegistos, besides the Everyday Writing website.

Figure 9 illustrates a 'univariate' search, that is, a search that involves a single variable. Our tool also accommodates 'bivariate' searches, entailing the examination of the relationship between two variables. The capabilities of this search type are illustrated in Figure 10 below, which shows a search for archives in the Everyday Writing corpus in relation to two periods, the Roman (I-III AD) and Late Antique (IV-VIII AD) period, in the form of a heatmap visualization, a graphical representation of data where values in a matrix are represented as colors. The numbers that are shown in this visualization refer to the number of texts from the Everyday Writing corpus that are included in the different archives.

Fig. 9: Univariate search (archives).

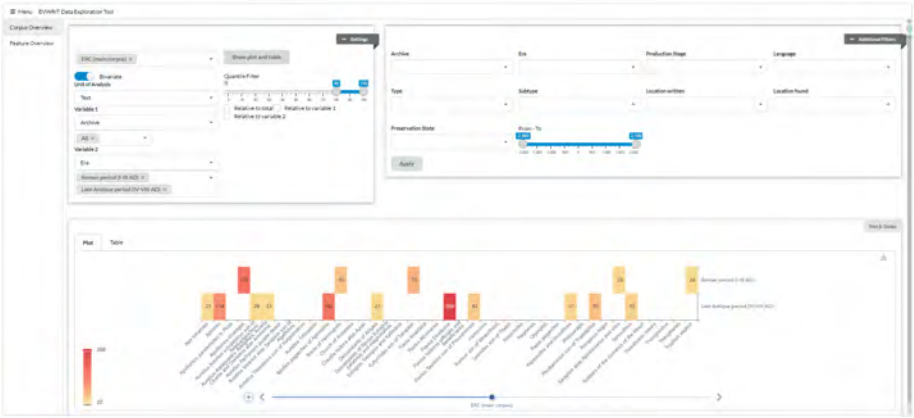


Fig. 10: Bivariate search (archives in relation to period).

The data exploration tool has a number of features that greatly facilitate searching for and displaying quantitative information of this type: as one can see, both for corpus and feature overview, there are ample additional filters, so that one can choose to only display information pertaining to a specific text type, location, era, language, date range, etc. One can also filter the information that is displayed in the types of visualization shown above by only displaying values that fall within a certain quantile range; for example, in both Figure 9 and 10, the quantile range has been set to 80-100, which means that only the top 20% is taken into account, or, in other words, only the 20% most frequently occurring archives are included in the visualization. Additionally, one can choose to only display in the visualization numerical information that falls between a certain frequency range. In Figure 10, for example, apart from the quantile filter that has been applied, as a result of which only a select number of archives is displayed on the X-axis, only values are shown for those archives which contain twenty texts or more.

Another important feature that the application has is that the user can normalize the frequencies that are given, so that percentages are displayed, in relation to the horizontal axis, the vertical axis, or the total. Figure 11 shows the percentual weight of our archives for each of the two periods considered (Roman vs. Late Antique), indicating, among others, that the archive of Apollonios the *strategos* covers 38.12% of the texts in the Everyday Writing Corpus in the Roman period.

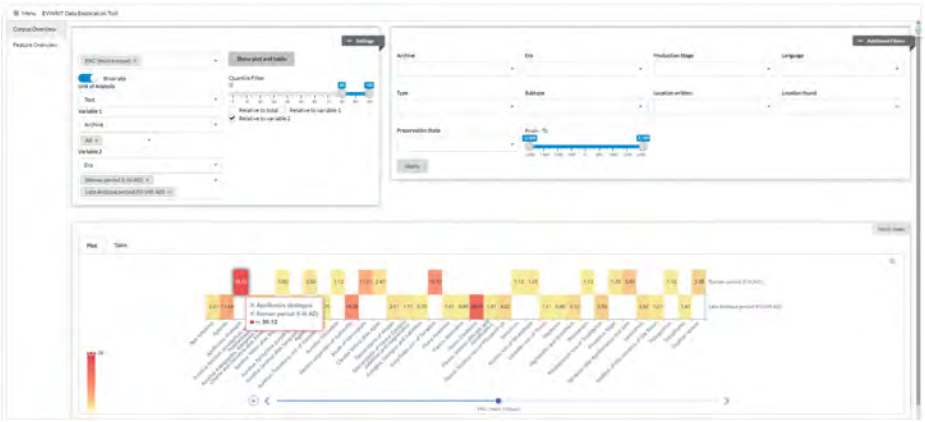


Fig. 11: Normalized bivariate search (archives in relation to period).

Another interesting feature that is worth mentioning here is that one can simultaneously display information for different research projects (as defined in the FileMaker database), and compare visualizations for these (sub)corpora, either manually by clicking the left and right arrows to switch between them, or automatically by clicking the 'play' button.

Given the amount of information in the Everyday Writing database, visualizations such as the ones shown here represent an essential tool to engage in data exploration, both of metadata and materiality and more specific features, and to explore potential patterns or links between types of data. In what follows, I will briefly illustrate the types of analysis that one can engage in with the help of both the Everyday Writing data exploration tool and the Everyday Writing website.

5 Two test cases: engaging in socio-semiotic, multimodal analysis

In an important book about the fundamentally social nature of human communication,[67] the linguistic anthropologist Michael Silverstein distinguishes between two processes that he considers central to human communication, which are called 'entextualization', (defined as 'the process of coming to textual formedness')[68] and 'contextualization' (defined as 'the process of how discourse points to (indexes) the context which seems to frame it',[69] that is, how interactants socially position themselves towards each other through the gradual progression of discourse). These two processes correspond to different types of organization in discursive events, namely *denotational text(uality)* (referring to the emergent coherence of what has been and will be said) and *interactional text(uality)* (the emergent coherence of what has been and will be done in terms of social action). As Silverstein points out, these two kinds of meaningfulness are intimately related to each other, standing as they do in a dialectical relationship: 'how you say what-you-say about whatever or whomever you're communicating about, comes to count interactionally as what-you-do in the way of creating the social organization of an ongoing interaction with a communicating other'.[70]

While the Everyday Writing project may seem rather open-ended in terms of the communicative features that are being annotated in order to get a better grasp of ancient interactional textuality, it has, in fact, a rather narrow focus, as it focuses specifically on discourse-organizational aspects of the text, which Michael Silverstein refers to in terms of its 'metricalization'. In a recent publication,[71] I have tried to systematize the features that we study in the Everyday Writing project by making reference to discourse 'frames' that are situated at three hierarchical levels (micro-, meso- and macro-), and that pertain to two different 'modes' of communication (with a basic distinction be-

67 Silverstein 2023.
68 Silverstein 2019, 56.
69 Silverstein 2019, 56.
70 Silverstein 2014, 499.
71 Bentein 2023b.

tween a 'linguistic' and 'visual' or 'typographical' mode of meaning making), as shown in Figure 12.

	Language	Framing features	Typography	Framing features
Micro-level framing	Clause/sentence	E.g., particles, subordinating conjunctions	Line	E.g., line fillers, word splitting, enlargement of letters
Meso-level framing	Thematic unit	E.g., particles, formulaic phrases	Lay-out unit	E.g., blank space, alignment, lectional signs
Macro-level framing	Generic part	E.g., formulaic phrases	Lay-out part	E.g., blank space, alignment, lectional signs, indentation
	Text	E.g., formulaic phrases	Page	E.g., margins, material substrate

Fig. 12: Multi-modal discourse segmentation (from Bentein 2023b, 93, Table 7.1).

As I mention in the same chapter, there are undoubtedly other features to be considered (e.g. handwriting as a visual type of meso- or macro-level framing, orthography as a linguistic/visual type of micro-level framing etc.), as well as, perhaps, other modes of meaning making. Rather than focusing on that discussion here, I want to briefly illustrate the types of research one can engage in with the digital tools that we have created in the context of the Everyday Writing project by making a distinction between two types of analysis, corresponding to two different perspectives, one top-down or 'macro-sociological' (§5.1), which involves creating what I like to call a 'semiotic grammar', and the other bottom-up or 'discourse-analytical' (§5.2), which involves engaging in what I like to call 'semiotic discourse analysis'. Evidently, in actual practice, these two perspectives can, and often are, combined, but for clarity's sake I will try to keep them apart here.

5.1 The macro-sociological perspective:[72] Semiotic grammar

In his 1989 contribution about the 'ethnography of writing', the aforementioned cultural and linguistic anthropologist Keith Basso programmatically describes the sort of research that he envisions, arguing that not only an adequate code description is needed, but that one also needs to turn one's attention to the code's manipulation in specific

72 I refer to the first perspective as 'macro-sociological', even if not all of the social variables in the database are necessarily situated at the macro-level (see Bentein 2019a, 131–4 for further discussion of contextual levels).

communicative settings: 'what is called for, essentially, is a grammar of rules for code use together with a description of the types of social contexts in which particular rules (or rule subsets) are selected and deemed appropriate.'[73] I refer to this of 'grammar' of how communicative choice relates to contextual variables in terms of 'semiotic grammar', a term that I borrow from an older publication by William B. McGregor,[74] which, however, focuses solely on language.

Let me briefly illustrate this perspective with a discussion of the social meaning-making potential of the visual and material features of documentary texts. That different handwriting styles – ranging from what we call in the project 'professional' and 'cultured' to 'graphic semi-illiterates' – are suited to different contexts of writing in terms of formality and/or the social status of the initiator and receiver is perhaps relatively self-evident, but scholarship has also drawn attention to other types of correlations between the physical characteristics of documentary sources in terms of their material composition, format (size, shape and orientation) and layout, and their creation and use within a specific socio-cultural context. As I mentioned in the introduction to this chapter, this point has been most explicitly made by Jean-Luc Fournet under the heading of what he refers to as "paléographie signifiante".[75] Fournet focuses in particular on aspects of documents' format, such as their *writing direction* and *shape*, which, he argues, are intimately connected to the documents' intended purpose. The validity of Fournet's point has been brought out by the Geneva-based *Grammateus* Project, which provides a typology of different documentary text types, not only in terms of their generic structure, but also their format and layout.

From a social-semiotic perspective, aspects such as documents' writing direction and shape can be seen as 'variables' with two or more variants, e.g. *perfibral* vs. *transfibral* for writing direction, or *horizontal* vs. *vertical* for orientation. Michael Silverstein refers to such variation in terms of a 'pragmatic paradigm', referring to the fact that during communication speakers can often choose one of a set of variant forms, each of which carries specific social indexicalities, being linked to a social situation of a particular kind. A typical example of such a pragmatic paradigm are different forms of address which one can use depending on the situation (e.g. 'dear Sir' vs. 'hey brother'), but in principle formal alternants need not be limited to the linguistic domain. From this perspective, most of the annotation fields in the Everyday Writing database can be seen as forming pragmatic paradigms, often containing an extensive set of formal alternants, particularly in the linguistic domain. One characteristic that arguably sets apart the visual/material domain is the existence of non-discrete, numerical variables in specific subdomains such as *height* and *width*, *margin size* (top, bottom, left, right), *line height*,

73 Basso 1989, 428.

74 McGregor 1997. Also note the concept of 'communicative' grammar by Leech – Svartvik 2002.

75 Understanding 'paleography' in a very broad sense, including the study of scripts, writing supports, formats, and layouts.

and *interlinear space height*. Such information is not only difficult and time-consuming to collect, but its non-discrete nature also poses a challenge to the human mind, as it is much more difficult to securely connect to specific contextual variables.

With the help of the Everyday Writing measurement tool, which I briefly mentioned in §3.2 and which is described more elaborately in Serena Causo's contribution to this volume, the Everyday Writing team has been able to collect a substantial amount of data with regard to the shape of documents. There are, of course, still severe limitations to this type of work, which are related both to the nature of the source material and the current state of our digital tool: at this point, we have mainly used the measurement tool for documents that are in a good state of preservation, that have good-quality imaging, that are written in a single column, that do not consist of multiple 'levels', etc. As I explained under §3.2, the measurement tool is based on bounding box annotation, the process of marking objects in images with rectangular shapes (bounding boxes) to identify and locate them. In many cases, however, the document in its entirety does not have a rectangular shape (which is of course true for papyri but also for ostraca), lines do not run straight, etc. That being said, we can, in preliminary fashion, draw attention to some tendencies in the Everyday Writing corpus, with the help of our Everyday Writing data exploration tool.

We have collected information on the variables *width* and *height* for nearly all of the documents in the Everyday Writing corpus, either on the basis of the information provided by the measurement tool or, when not possible, by the text editions. Restricting ourselves to documents that are not broken on the top/bottom for height, or to the left/right for width, we can now use the Everyday Writing data exploration tool to display the distribution of documents in our corpus in terms of height and width. Figure 13 plots the height range of documents in our corpus, only displaying results that fall within the 80-100 percentile range.

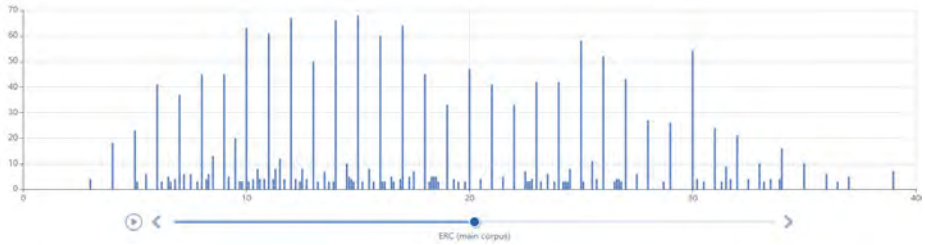


Fig. 13: Height of documents in the Everyday Writing corpus

Further analysis of the data exported through the data exploration tool indicates that the texts that are complete in terms of one or both of the dimensions, have an average height of approximately 23.27 cm, and an average width of approximately 16.78 cm.

Bivariate analysis of height and width in correlation with other metadata variables allows us to greatly diversify these averages, however. In terms of historical period, the average height and width are higher in the Late Antique period than they are in the Roman period, for example (average H24.46 vs. 21.04 cm., W19.64 vs. 13.24 cm.). Between different text types, too, there are some noticeable differences: in the Roman period, for example, the mean/median width is similar for letters, petitions, and contracts, but there are striking differences in terms of height, the mean/median height of contracts being significantly higher than letters, and that of petitions being slightly higher than contracts, as shown in Figure 14.[76]

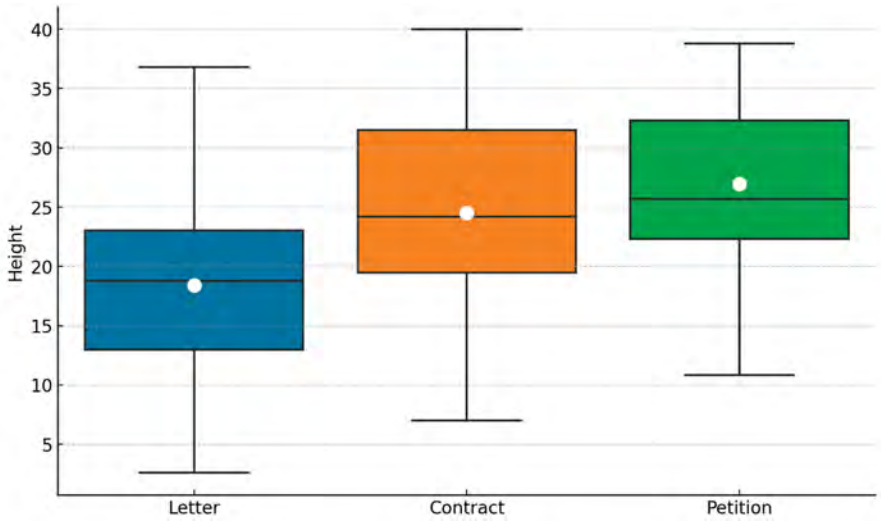


Fig. 14: Boxplot of height for macro-categories of texts (Roman period)

The data that we have gathered also allows us to detect variation inside these larger generic categories: letters and contracts, unsurprisingly, are most varied. Limiting ourselves to the Roman period, for example, private and especially invitation letters are the smallest in terms of width and height, whereas recommendation and official letters are the largest. At this point, the data exploration tool allows us to visualize this variation by using the filters and then creating separate univariate visualizations for each text type; in the future, we would like to make it possible to create a bivariate heatmap visualization whereby non-discrete values are 'binned' in user-defined groups. Visualization also shows the degree of variation inside specific categories: invitation letters are

76 In this image, the white dot marks the mean and the horizontal line the mean. Statistical processing of the type found in this figure is currently done outside of the Everyday Writing data exploration tool.

relatively inform in terms of height, for example, which is much less true for official letters. To better understand the nature of these differences, it would be interesting to relate height to other materiality categories, such as the total number of words and lines, but I will not go further into this here.[77]

5.2 The discourse-analytical perspective: Semiotic discourse analysis

The second perspective that I want to explore here is oriented bottom-up rather than top-down, in that its main purpose is to better understand specific texts, in particular how interactants socially position themselves towards their addressees by bringing together different types of communicative features that together 'shape' or 'frame' the textual message.[78] This naturally entails paying more attention to features that form 'outliers' from a macro-sociological point of view, and moving away from the idea that there exists a strict connection between one specific communicative feature and one specific social value[79] (e.g. 'sir' signaling 'formal' in English). Instead, one can think of the social meaning that is signaled by communicative features in terms of a field of associated values – an 'indexical field' – that can be dynamically and strategically employed to create a socially multi-layered message.[80]

This type of dynamic qualitative perspective is not uncommon in (historical) sociolinguistics – in particular in (sub)fields such as interactional sociolinguistics and conversation analysis – but has been less often explored in a multi-modal way. In what follows I want to briefly discuss how it can be pursued through means of the digital tools that we have created, by engaging with a linguistic domain that I have studied relatively intensively in the last few years, namely subordination, in particular verbal complementation. In a series of articles that I published, I argued that the choice for both 'minor' and 'major' complementation patterns seems to be governed by sociolinguistic (pragmatic), rather than semantic factors. The choice for ὅτι vs. ὡς after communication verbs, for example, seems to depend less from the concept of 'factivity' (whether or not the truth value of the complement is presupposed), as it was in the Classical period,[81] but rather on the context of writing, including such aspects as the formality of

77 For the importance of considering size in a relative fashion, see Stroppa 2023, 29.

78 For the relationship between framing and semiosis, see e.g. Bentein 2023b, 89–90.

79 A view that is still maintained in sociolinguistic studies, implicitly or explicitly. See Bentein 2019a, 145–6.

80 For further discussion of the concept of indexicality, see Bentein 2019a.

81 E.g. van Emde Boas – Rijksbaron – Huitink – de Bakker 2019, 504–5, "in classical Attic ὅτι is the default conjunction … ὡς is mostly used if the reporter expressly wishes to convey that the truth of the reported statement is open to doubt."

the communication, the hierarchical or symmetrical relationship between the initiator and the receiver, their respective social ranks, etc.[82]

An element that was not incorporated in these studies, but which has been annotated systematically in the context of the Everyday Writing project, is the syntagmatic ordering of the matrix verb and the complement clause, that is, whether the matrix verb precedes or follows the complement clause. Ancient Greek constituent order is, of course, an immense topic that I cannot engage with in any detail in the context of the present contribution: suffice it to say that already in the Classical period it was uncommon for complement clauses to precede the matrix verb,[83] which was also, and arguably even more so, the case in the Post-classical period,[84] when VSO/SVO order, rather than SOV order, became standard.[85] This is not a unifying trend, however: one does still find preposed complement clauses,[86] particularly with formulaic non-finite complement clauses such as the disclosure formula γινώσκειν σε θέλω "I want you to know" and the farewell greeting ἐρρῶσθαί σε εὔχομαι "I bid you farewell" in informal texts, and the concluding acknowledgement formula ταῦθ' οὕτως ἔχειν δώσειν ποιεῖν φυλάττειν ὡμολώγησα "I agreed that the things are so, and so to give, do and keep" in formal texts. Figure 15 shows the number of pre-posed vs. post-posed complement clauses, in relation to their formulaicity.

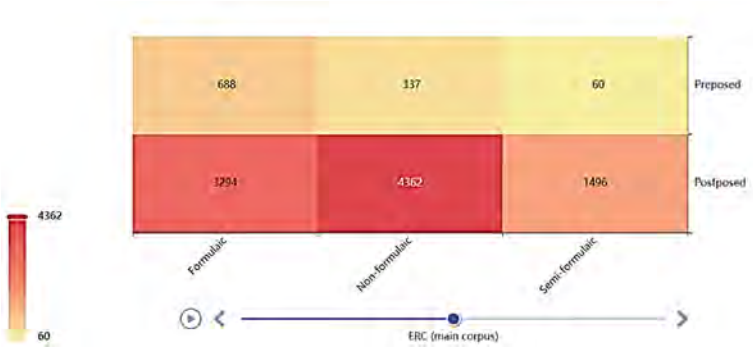


Fig. 15: Pre- vs. postposition in relation to formulaicity.

82 For a larger-scale investigation of complementation patterns in documentary papyri, see now Keersmaekers 2020.

83 Compare e.g. Allan 2012, 11 for the order of main verb and infinitive in classical prose.

84 Turner 1970, 344–5 observes for the New Testament that "normally the dependent clause follows the main clause", though his discussion mostly focuses on adverbial clauses.

85 See e.g. Levinsohn 2000, 16–7; Horrocks 2007, 620–3; Kirk 2012.

86 In the database, we also annotate a third type of word order pattern, called 'scrambled' (compare Allan's notion of "clause intertwining", Allan 2012), which I will not go further into here.

When we concentrate solely on non-formulaic complement clauses (which, to a large extent, are non-finite), and relate them to the social distance values 'formal' vs. 'informal',[87] we find that a significant correlation seems to exist between formality and the pre- vs. post-positioning of the complement clause (with up to 63% of the preposed complement clauses occurring in formal texts), which also becomes apparent from close reading of formal documents such as contracts and petitions.[88] This does not come as a major surprise, since there seems to be a natural correlation between formality and archaizing language.[89]

This does raise the question which sorts of informal texts employ preposed complement clauses and whether or not this is purposely done; this may, in turn, suggest the need for further refinement of the binary formality scale that we are currently employing in the database (which is based on the distinction between 'official' and 'non-official' documents), in the sense that further subcategories need to be recognized. In order to better understand the motivation behind the choice for preposed complement clauses, one can pursue a number of research strategies, such as analyzing the extent to which writers do or do not vary between pre- and postposed complement clauses inside one and the same text, and/or analyzing the co-occurrence of preposed complement clauses with some of the other types of linguistic and non-linguistic framing features[90] that I listed in our previous Figure 12.

These and other questions belong to the field of 'semiotic discourse analysis' properly speaking. The important point to note here is that the digital tools that we have developed and that I have outlined under §4 allow us to pursue such questions. The Everyday Writing website, for example, allows one to sort texts on the basis of the (relative) frequency of occurrence of preposed complement clauses, which provides a key entry point to our current research question. Obviously it goes beyond the scope of this contribution to engage in a full-blown discourse analysis of the social semiotics of sentential syntax,[91] but it is worth briefly discussing one text which strongly suggests that, indeed, preposition of complement clauses can be strategically employed (manipulated) as one of

87 See Bentein 2017, 22 for discussion of the relationship between social distance and formality. For the key importance of formality in explaining linguistic choice, see John Lee's landmark paper on the Greek New Testament (Lee 1985).

88 By way of illustration, see e.g. the contract of will P.Oxy. XXVII 2474 (III AD) = TM 30460, in which provisions are made through preposed complement clauses (ἔχειν ᾳ[ὐτὴ]ν θέλω, l. 20; ἐλευθέρους εἶναι θέλω, l. 28; τήν γε ὁμογνησίαν μου ἀδελφὴν Θεογνώστην παραμεῖναι θέλω, ll. 31–32) etc.

89 For the connection between archaism and formality, see Bentein 2019a, 154–5. In Bentein 2017, I show that the use of non-finite complementation compared to finite complementation, too, can be related to a number of social variables, including social distance.

90 There is, of course, no need to restrict oneself to framing features. It would be worth, for example, looking at head – dependent structures at the nominal level, too, for example (following John A. Hawkin's concept of 'cross-category harmony', Hawkins 1982).

91 In Bentein 2020 I further discuss the cognitive salience and associated sociolinguistic sensitivity of syntactic features compared to features situated at other linguistic levels.

a number of co-occurring features contributing to self-positioning and interpersonal alignment.[92] Such co-occurrence is in line with findings in sociolinguistics and linguistic anthropology, which suggest that in the case of non-referential indexical features[93] such as word order patterns, social entailments are 'less the effect of the particular "salient" or "overt" sign in question (e.g., a pronoun) than the total effect of a textual configuration of indexical signs (e.g., the pronoun, previous/subsequent address practices, bodily hexis, etc.).'[94] In other words, it is the *totality* of signs that is employed and their co-occurrence that guides the continually ongoing process of contextualization.

P.Oxy. I 122 = TM 31348, the document that I briefly would like to discuss here,[95] is a rather short letter dated to the late third/fourth century, which is somewhat ambiguous in terms of its formality, as we will see below. The letter is sent by Gaianus to Agenor, the latter being referred to in the external address on the verso as 'prefect', to be understood as prefect of a legion, according to the editors. In this fifteen-line letter, we find two instances of preposed complement clauses, one formulaic in the closing, and another non-formulaic in the letter's body: ἀγρεύειν τῶν θηρίων δυνά[με]θ̣α οὐδὲ ἕν "we cannot catch a single animal". It seems that the letter writer consciously made an effort to vary his word order here, since the passage contains two examples of *dependent – head* structures – one clausal, ἀγρεύειν … δυνά[με]θ̣α, and one nominal τῶν θηρίων … οὐδὲ ἕν – which are discontiguous and intertwined, perhaps for reasons of emphasis (with emphasis on ἀγρεύειν and οὐδὲ ἕν). Otherwise, too, the letter is quite interesting in terms of textual composition and interpersonal positioning, showing a high degree of what I have elsewhere referred to as 'discourse planning' and which can be defined as the amount of attention that is paid to the process of textualization.[96] With the help of the Everyday Writing website, the analyst can better understand the textual makeup of the letter, and, by extension, how the writer positions himself/his letter in the textual landscape.[97]

92 This degree of co-occurrence is sometimes referred to in terms of 'intersemiotic complementarity' (e.g. Royce 2007).

93 For different types of indexicality, see Bentein 2019a.

94 Nakassis 2018, 294.

95 Extensive discussion of the language of this text is also offered by Luiselli 1999, 227–32, who pays attention to the unusual word order of the complement clause, as well as a number of other linguistic features.

96 On planned and unplanned discourse, see further Ochs 1979.

97 In what follows, I use the descriptive framework outlined in Bentein 2023b, moving from the macro-level to the micro-level (compare, in particular, the qualitative discussion in Bentein 2023b, 101–5).

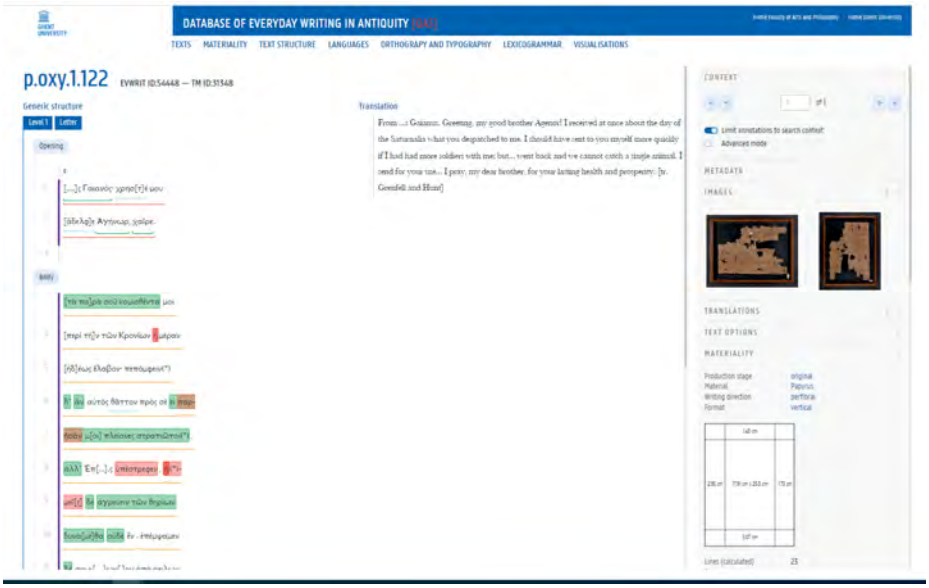


Fig. 16: Website representation of P.Oxy. I 122 (III-IV AD) = TM 31348.

As the purple line to the left-hand side of the text indicates, from a macro-level point of view, the letter is written in a professional, right-leaning hand, which, according to the editors at least, has a 'thoroughly Latin' appearance.[98] The letter has ample margins, as indicated in the information sidebar under 'materiality', including a straight right margin of 1.75 cm., a feature that is rather unusual in letters, and even non-literary documents in general.[99] Its opening and closing lines are visually separated from the body, with a large vertical space following the opening, and the closing being written in a distinct type of handwriting (more right-leaning and cursive than the body), and right-aligned. From a generic point of view, the document is framed by an epistolary opening and closing, which only follow conventions partly: instead of the more usual pattern nominative + dative + χαίρειν, the imperatival form χαῖρε is used, with the name of the initiator in the nominative and that of the receiver in the vocative case.[100] The closing

98 'That Gaianus was more accustomed to Latin than Greek is very evident from his handwriting, which is marked by a thoroughly Latin appearance throughout, and by an occasional obtrusion of Latin forms of letters, e.g. m.' (Grenfell – Hunt 1898, 189).

99 Of a test set of 1097 completely preserved documents, only 3.92% of the documents had a right margin of 1.75 cm or higher.

100 See Fournet 2009, 37–46 for the transformation of the letter opening in Late Antiquity; Fournet considers the introduction of χαῖρε or χαίροις as an intermediate step ultimately leading to the less of the prescript (Fournet 2009, 45). Unfortunately, we do not have a very precise date for our letter, so that it is difficult to position our letter in this new trend.

greeting follows the standard pattern ἐρρῶσθαί σε εὔχομαι, but with the addition of the health verb προκόπτειν "I prosper". I have discussed the use of such innovative features, and how they are used for social positioning, elsewhere, so I will not go further into this here.[101] Rather, I would like to draw attention to the fact that at the meso- and micro-level, too, our document shows a high degree of discourse planning, with the use of the particle δέ signaling the different segments in the body of the text (all of which are related to the sending and receiving of goods), which is visually matched by the insertion of small horizontal spaces, another, relatively uncommon[102] feature at a time when *scriptio continua* was standard practice.[103] This double segmentation (linguistic and visual) can be seen in our generic and lay-out visualizations (with only the generic visualization being displayed in Figure 16). What is even more unusual is that at the level of individual words, our writer has included a spiritus asper for three words starting with a long aspirated vowel (ἡμέραν, l. 4; ἡμεῖ[ς], ll. 8–9; ᾧ, l. 12), as well as a *diaeresis* on the initial vowel of the verb ὑπέστρεφεν, a feature that is more common in the papyrological corpus than is the use of a spiritus asper.[104] The use of a preposed complement clause, too, forms, as we have seen, a relatively unusual feature at the micro-level.

What is interesting is that the high amount of attention that is paid to discourse planning is not immediately matched by the contents of the letter, or the use of non-referential indices such as forms of address,[105] resulting in complex social positioning. The editors themselves comment that the remarks that are made in the body of the letter have a familiar tone, and observe that the initiator makes no effort to place his name after that of the receiver, which represented a common politeness procedure.[106] Despite the fact that the receiver is identified as a 'prefect', he is addressed as 'my dear brother' in the opening, and as 'my lord brother' in the closing, which again signal familiarity, and according to the editors, suggests that the initiator must have been of high rank, too, as confirmed by the mentioning of soldiers under his command in the body of the text.[107] Semiotic discourse analysis of the type I have engaged in here shows how writers

101 Bentein 2023a.

102 Though less uncommon than sometimes thought. See e.g. Bentein 2023b, 95 for horizontal spaces in the corpus of women's letters.

103 The letter also has some small spaces between words and word groups e.g. at line 6 between θᾶττον and πρὸς σὲ, but I will not go further into this here.

104 At present, only 38 unique texts in the Everyday Writing corpus have been automatically annotated for one or more asper signs, 29 of them from the Late Antique period. Fournet 2009, 32–7 discusses the introduction of such diacritical signs in terms of a 'literarisation' of documentary practice. See also Fournet 2020.

105 For the (mixed) indexicality of forms of address, see Bentein 2019a, 149.

106 Sarri 2018, 42–3 notes that the names of the initiator and receiver are ordered according to their hierarchical relationship starting from the Ptolemaic period. Fournet 2009, 43 specifies that this was particularly the case in official letters, but that the practice was extended to private letter writing starting from the 3rd cent. AD, and was systematized in the 4th cent. AD.

107 Note that the document is classified as 'amtlich' in the DDbDP.

can strategically manipulate different types of communicative features to create a complex, multi-layered representation signaling joint cultural background,[108] personal distinction (high status), and interpersonal closeness, among others. Further research should make clearer how common this mixture of features is;[109] it may suggest the need to recognize a formality scale that is more fine-grained than the current binary distinction between 'formal' and 'informal'.[110]

6 Concluding remarks

In this contribution, I have discussed the need to study everyday documents from a 'communicative' perspective, introducing the key notions of 'social semiotics' and 'multimodality'. After giving an overview of the two main digital tools that we have developed in the context of the Everyday Writing project, I related these tools to two distinct types of analysis, which I referred to as 'macro-sociological' (*semiotic grammar*) and 'discourse-analytical' (*semiotic discourse analysis*). While these two perspectives may seem quite disparate, the first being quantitative and top-down, and the second qualitative and bottom-up, in reality they are closely connected to each other. This is also reflected in the way we use our tools: while utilizing the data exploration tool, it is common to encounter outliers[111] that necessitate further analysis on the website. Conversely, while working with the website, one often encounters features that appear uncommon, prompting the utilization of the data exploration tool to gain a better understanding of their level of rarity.

It is important to emphasize here that the digital infrastructure that we have developed in the context of the Everyday Writing project is in no way meant to replace existing tools/portals such as the *DDbDP*, *Trismegistos* or *Grammateus*. Our tools have been developed for a small focus corpus and have a specific research orientation. For the short-term future, our plans are to finish the manual annotation process (which, unfortunately, is time-consuming and error-prone); to improve export functionalities for our dataset, as well as linking to other digital portals; to create an online documentation explaining our data structure and defining/illustrating the values that we have chosen

108 It is interesting to note that in Latin, too, there is a tendency towards postposition of the infinitive in the type of complement structure found in our letter, preposition seemingly being used for reasons of emphasis (further discussion in Pinkster 2021, 1126–9). As such, one could hypothesize that introduction of a structure such as ἀγρεύειν τῶν θηρίων δυνά[με]θα οὐδὲ ἕν in a Latinate context might have a more complex social effect than discussed above. Likewise, the use of the opening greeting χαῖρε + vocative could be seen to mimick the Latin epistolary opening *salve* + vocative (Sarri 2018, 49).

109 Sarri 2018, 67–8 discusses the 'friendly' character of official letters written between officials at an equal administrative level.

110 For more finegrained typologies of formality, see e.g. Joos 1967; Hall 1990.

111 E.g. values that fall outside the interquartile range.

for specific fields; and to more systematically describe, at least for a select number of variables, their 'semiotic potential', that is, how they relate to metadata fields such as *social distance*, *social rank*, *era*, *provenance*, etc.

The availability of funding for the further development of our digital tools remains uncertain, and only time will reveal its outcome. There is, in any case, a lot of potential for further improvement: it would be beneficial, for example, to integrate to a greater extent the tools that we have now developed in different environments (FileMaker, MySQL, RShiny); to explore the potential of AI technology in expediting the annotation process; to find a more flexible way to include different types of 'base annotations' (§3.4), which unlike the other types of annotations discussed in §3 are potentially open-ended; to extend the quantitative methods that we apply to the dataset so as to find clusters between features and metadata categories; etc., to name but some desiderata. As our database has been used by researchers working on different cultural traditions, including Greek, Coptic, Latin, and Arabic documentary sources, it would be worth engaging more explicitly in cross-cultural comparison than we have done so far, and to develop the digital technology to do so.[112]

Bibliography

Allan, R. J. (2012), *Clause Intertwining and Word Order in Ancient Greek*, Journal of Greek Linguistics 12, 5-28.

Amory, Y. (forthcoming), *More than a Simple Intuition. Towards a Categorization of Paleographical Features in Greek Documentary Papyri*, Comparative Oriental Manuscript Studies Bulletin.

Basso, K. H. (1989), *The Ethnography of Writing*, in *Explorations in the Ethnography of Speaking*, ed. by J. Sherzer – R. Bauman, 2nd ed., Cambridge, 425–32.

Bateman, J. A. (2008), *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*, Basingstoke – New York.

Bateman, J. A. – J. Wildfeuer – T. Hiippala (2017), *Multimodality: Foundations, Research and Analysis. A Problem-Oriented Introduction*, Berlin.

Bentein, K. (2015a), *Minor Complementation Patterns in Post-Classical Greek (I-VI AD): A Socio-Historical Analysis of a Corpus of Documentary Papyri*, Symbolae Osloenses 89, 104–47.

Bentein, K. (2015b), *Particle-Usage in Documentary Papyri (I-IV A.D.): An Integrated Sociolinguistically-Informed Approach*, Greek, Roman and Byzantine Studies 55, 721–53.

Bentein, K. (2017), *Finite vs. Non-Finite Complementation in Post-Classical and Early Byzantine Greek*, Journal of Greek Linguistics 17, 3–36.

Bentein, K. (2019a), *Dimensions of Social Meaning in Post-classical Greek*, Journal of Greek Linguistics 19, 119–67.

Bentein, K. (2019b), *Historical Sociolinguistics: How and Why? Some Observations from Greek Documentary Papyri*, Annali dell'Istituto Universitario Orientale di Napoli 41, 145–54.

Bentein, K. (2020), *The Distinctiveness of Syntax for Varieties of Post-Classical and Byzantine Greek: Linguistic 'Upgrading' from the Third Century BCE to the Tenth Century CE*, in *Varieties of Post-Classical and Byzantine Greek*, ed. by K. Bentein – M. Janse, Berlin – New York, 381–414.

112 See Bentein – Kootstra forthcoming for a pilot study.

Bentein, K. (2023a), *A Typology of Variations in the Ancient Greek Epistolary Frame (I-III AD)*, in *Historical Linguistics and Classical Philology*, ed. by G. Giannakis – E. Crespo – J. de La Villa – P. Filos, Berlin, 415–57.

Bentein, K. (2023b), *The Textualization of Women's Letters from Roman Egypt. Analyzing Historical Framing Practices from a Multi-Modal Point of View*, in *Novel Perspectives on Communication Practices in Antiquity. Towards a Historical Social-Semiotic Approach*, ed. by K. Bentein – Y. Amory, Leiden, 89–112.

Bentein, K. – Amory, Y. (2023), eds., *Novel Perspectives on Communication Practices in Antiquity. Towards a Historical Social-Semiotic Approach*, Leiden.

Bentein, K. – Kootstra, F. (forthcoming), *Visually Structuring Text: A Comparative Multimodal Analysis of the Greek and Arabic Qurra Papyri*, Segno & Testo.

Bernhart, W. – Wolf, W. (2006), *Framing Borders in Literature and Other Media*, Amsterdam – New York.

Buijs, M. (2005), *Clause Combining in Ancient Greek Narrative Discourse: The Distribution of Subclauses and Participial Clauses in Xenophon's* Hellenica *and* Anabasis, Leiden – Boston.

Carlig, N. (2020), *Les symboles chrétiens dans les papyrus littéraires et documentaires grecs : forme, disposition et fonction (IIIe – VIIe/VIIIe siècles)*, in *Signes dans les textes. Continuités et ruptures des pratiques scribales en Égypte pharaonique, gréco-romaine et byzantine*, éd. par N. Carlig – G. Lescuyer – A. Motte – N. Sojic, Liège, 271–81.

Crystal, D. (1979), *Reading, Grammar and the Line*, in *Growth in Reading*, ed. by D. Thackray, London, 26–38.

Fournet, J.-L. (2007), *Disposition et réalisation graphique des lettres et des pétitions protobyzantines: pour une paléographie 'signifiante' des papyrus documentaires*, in *Proceedings of the 24th International Congress of Papyrology, Helsinki, 1-7 August, 2004*, ed. by J. Frösén – T. Purola – E. Salmenkivi, Helsinki, I, 353–67.

Fournet, J.-L. (2009), *Esquisse d'une anatomie de la lettre antique tardive d'après les papyrus*, in *Correspondances. Documents pour l'histoire de l'Antiquité tardive. Actes du colloque international, Université Charles-de-Gaulle-Lille 3, 20-22 novembre 2003*, éd. par R. Delmaire – J. Desmulliez – P.-L. Gatier, Lyon, 23–66.

Fournet, J.-L. (2020), *Les signes diacritiques dans les papyrus documentaires grecs*, in *Signes dans les textes. Continuités et les ruptures des pratiques scribales en Égypte pharaonique, gréco-romaine et byzantine*, éd. par N. Carlig – G. Lescuyer – A. Motte – N. Sojic, Liège, 145–66.

Grenfell, B. P. – Hunt, A. S. (1898), eds., *The Oxyrhynchus Papyri,* I, London.

Hall, E. T. (1990), *The Hidden Dimension*, New York.

Halliday, M. A. K. (2010), *Language and Social Man*, in *Language and Society. Volume 10 in the Collected Works of M.A.K. Halliday*, ed. by J. Webster, London – New York, 65–130.

Halliday, M. A. K. – Hasan, R. (1989), *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*, 2nd ed., Oxford – New York.

Halliday, M. A.K. –Matthiessen, C. M. I. M. (2013), *Halliday's Introduction to Functional Grammar*, London – New York.

Hawkins, J. A. (1982), *Cross-Category Harmony, X-Bar and the Predictions of Markedness*, Journal of Linguistics 18, 1–35.

Hiippala, T. (2016), *The Structure of Multimodal Documents: An Empirical Approach*, New York.

Horrocks, G. C. (2007), *Syntax: From Classical Greek to the Koine*, in *A History of Ancient Greek*, ed. by A.-P. Christidēs, Cambridge, 618–31.

Joos, M. (1967), *The Five Clocks*, New York.

Keersmaekers, A. (2020), *A Computational Approach to the Greek Papyri. Developing a Corpus to Study Variation and Change in the Post-Classical Greek Complementation System*, PhD Diss., K. U. Leuven.

Kirk, A. (2012), *Word Order and Information Structure in New Testament Greek*, PhD Diss., Leiden University.

Kress, G. R. – Van Leeuwen, T. (1996), *Reading Images: The Grammar of Visual Design*, London – New York.

Lee, J. A. L. (1985), *Some Features of the Speech of Jesus in Mark's Gospel*, Novum Testamentum 27, 1–26.

Leech, G. N. – Svartvik, J. (2002), *A Communicative Grammar of English*, 3rd ed., Harlow.

Lemke, J. (1998), *Multiplying Meaning: Visual and Verbal Semiotics in Scientific Text*, in *Reading Science: Critical and Functional Perspectives on Discourses of Science*, ed. by J. R. Martin – R. Veel, London, 87–113.

Levinsohn, S. H. (2000), *Discourse Features of New Testament Greek: A Coursebook on the Information Structure of New Testament Greek*, 2nd ed., Dallas.

Lillis, T. M. (2013), *The Sociolinguistics of Writing*, Edinburgh.

Logozzo, F. (2015), *Register Variation and Personal Interaction in the Zenon Archive*, Studi e Saggi Linguistici 53, 227–44.

Luiselli, R. (1999), *A Study of High Level Greek in the Non-Literary Papyri from Roman and Byzantine Egypt*, PhD Diss., University of London.

Luiselli, R. (2008), *Greek Letters on Papyrus, First to Eighth Centuries : A Survey*, Asiatische Studien 62, 677–737.

McGregor, W. (1997), *Semiotic Grammar*, Oxford – New York.

Nakassis, C. V. (2018), *Indexicality's Ambivalent Ground*, Signs and Society (Chicago, Ill.) 6, 281–304.

Ochs, E. (1979), *Planned and Unplanned Discourse*, in *Discourse and Syntax*, ed. by T. Givon, New York, 51–80.

Pinkster, H. (2021), *The Oxford Latin Syntax. Volume II, The Complex Sentence and Discourse*, 1st ed., Oxford – New York.

Ricceri, R. – Bentein, K. – Bernard, F. – Bronselaer, A. – De Paermentier, E. – De Potter, P. – De Tré, G. *et al.* (2023), *The Database of Byzantine Book Epigrams Project: Principles, Challenges, Opportunities*, Journal of Data Mining and Digital Humanities, https://hal.science/hal-03833929v3/file/DBBE_final.pdf.

Royce, T. D. (2007), *Intersemiotic Complementarity: A Framework for Multimodal Discourse Analysis*, in *New Directions in the Analysis of Multimodal Discourse*, ed. by W. Bowcher – T. D. Royce, Mahwah (NJ), 63–109.

Sarri, A. (2018), *Material Aspects of Letter Writing in the Graeco-Roman World: 500 BC--AD 300*, Berlin – Boston.

Sijpesteijn, P. M. (2013), *Shaping a Muslim State: The World of a Mid-Eighth-Century Egyptian Official*, Oxford.

Silverstein, M. (2014), *The Voice of Jacob: Entextualization, Contextualization, and Identity*, English Literary History 81, 483–520.

Silverstein, M. (2019), *Texts, Entextualized and Artifactualized: The Shapes of Discourse*, College English 82, 55–76.

Silverstein, M. (2023), *Language in Culture. Lectures on the Social Semiotics of Language*, Cambridge – New York.

Spitzmüller, J. (2013), *Graphische Variation als soziale Praxis: eine soziolinguistische Theorie skripturaler*, Berlin.

Stroppa, M. (2023), *BIG & Small: The Size of Documents as a Semiotic Resource for Graeco-Roman Egypt*, in *Novel Perspectives on Communication Practices in Antiquity: Towards a Historical Social-Semiotic Approach*, ed. by K. Bentein – Y. Amory, Leiden, 29–38.

Turner, N. (1970), *A Grammar of New Testament Greek: Volume III, Syntax*, 3rd ed., Edinburgh.

van Emde Boas, E. – Rijksbaron, A. – Huitink, L. – de Bakker, M. (2019), *Cambridge Grammar of Classical Greek*, Cambridge.

van Leeuwen, T. (2006), *Towards a Semiotics of Typography*, Information Design Journal 14, 139–55.

Walker, S. (2001), *Typography and Language in Everyday Life: Prescriptions and Practices*, Harlow – New York.

Serena Causo

Enhancing Data Collection on the Materiality of Papyri: The Measurement Tool

1 Introduction

The study of the materiality of ancient documents holds immense potential for advancing our understanding of historical documents and ancient writing practices. The importance of a consistent collection and analysis on the material aspects of written artefacts was already advocated by Turner in 1971, in his *Greek Manuscripts of the Ancient World*:

> an attempt at dating will begin by considering the material aspects of a manuscript: is the writing-material fine or coarse papyrus, skin or parchment? What is the size and format? Fashion apply to size and formats of manuscripts just as they do to hair-styles. A collection of the evidence is an urgent desideratum for palaeographers.[1]

The venture auspicated by Turner might have been of hard execution at the time in which he wrote it, when both inspecting the written objects and the access to images were no easy feats. In the last decades, on the other hand, the digitization of numerous collections by an ever-increasing number of institutions worldwide has created rapid access to high-quality images, supporting and boosting the interest of scholars towards the material aspects of documents.[2] Increasingly, researchers have started looking beyond the content, as a way of integrating and enhancing the information offered by the written texts through cues locked in the writing support.[3]

Noteworthy in the field of papyrology are the works dedicated to the analysis of the materiality of specific genres of texts, such as William Johnson's work devoted to the investigation of the characteristics of literary rolls and scribal practices in Oxyrhynchus, or Antonia Sarri's study on the material and visual aspects of private and official letters from the Ptolemaic and Roman period.[4] A recent collaborative volume edited by Nicola Reggiani explores the materiality of medical papyri.[5] More and more, scholars

1 Turner 1971, 22; 1978, 61.

2 The use of digital images as surrogates for the study of the materiality of artefacts comes with caveats, both for researchers in charge of studying and for archivists in charge of digitizing them. The main dangers are the de-materialisation, decontextualization and distortions of the material objects. For a discussion on these challenges, see Rekrut 2014.

3 Petrovic – Petrovic – Thomas 2019; Hoogendijk – van Gompel 2018; Angliker – Bultrighini 2023.

4 Sarri 2018.

5 Reggiani 2024.

https://doi.org/10.1515/9783111070162-015

engage in the analysis of the materiality of specific subtypes of documents, such as liturgical nominations,[6] cessions of cleruchic land,[7] warrants[8] and certificates of pagan sacrifices.[9] The importance of valuing ancient documents as cultural objects and the necessity of a broader investigation on their material and visual features was already advocated in 2007 by Jean-Luc Fournet, who coined the notion of "paléographie signifiante."[10] Since then, material and visual aspects of documents have been increasingly used as tools for exploring the various layers of meaning attached to written artefacts, emphasizing the interplay between the content and the container.[11] A stronger focus on materiality is also evident in the increased attention devoted to the description of the physical and visual aspects of the documents in the context of editorial practices.[12]

Still less frequently discussed, and yet very promising, are the studies on the materiality of ancient documents devoted to exploring the manufacturing process of the writing support and writing instruments, such as the type of ink. Numerous researches conducted by Myriam Krutzsch have shed light on aspects such as the quality, color or thickness of papyri, revealing significant developments of the manufacturing process and quality of the support.[13] Similarly, specific preferences in the choice of ink based on textual genres have emerged from the analyses conducted by Tea Ghigo and Alberto Nodar Dominguez.[14]

Significantly, the material turn has also motivated ethical discussions concerning the preservation, study and exchange of papyri, which must be regarded as historical objects and, as such, have been granted specific rights: the right to be curated and stored correctly or to be accessible and publicly available for inspection.[15]

For comprehensive and reliable studies on the materiality of ancient documents, the availability of quantitative data is paramount. As scholarly conversations on materiality intensify and more researchers incorporate the analysis of material aspects into their work, the need for databases that collect data on the material, visual, and typographical aspects of papyri and other writing supports becomes increasingly apparent.

In the context of the EVWRIT Project, we have annotated aspects of the materiality of approximately 6.000 objects between petitions, contracts, and letters from the Roman and Byzantine periods.[16] At the same time, the team of the CEDOPAL in Liège is currently

6 Schubert 2022.

7 Ferretti – Fogarty – Nury – Schubert 2020.

8 Schubert 2018.

9 Schubert 2016.

10 Fournet 2007 and 2022.

11 Fournet 2022; Stroppa 2022; Amory 2022. For a recent discussion on multi-modal and socio-semiotic approaches for the study of ancient documents, see Bentein – Amory 2022, 1–14.

12 Fournet 2019 and 2022.

13 See Krutzsch 2020; for further references on the subject, consult the bibliography included there.

14 Ghigo – Nodar 2023; for further references on the subject, consult the bibliography included there.

15 Mazza 2021, 376–93.

16 See the contribution by Klaas Bentein included in this volume.

developing a database for the materiality of ancient books, which collects codicological and palaeographical information on the entire corpus of Greek and Latin literary papyri.[17] These databases represent a first step towards collaborative research and data sharing among scholars, leading to a more interconnected and comprehensive study of ancient manuscripts.

Furthermore, in the past years, several projects have leveraged the advancements of Document Image Analysis (DIA)[18] and deep learning machines as an opportunity to apply modern technologies to the study of historical documents, with the scope of assisting, facilitating and expediting the work of researchers in the analysis of both physical and visual aspects of written objects.[19] In the field of papyrology, the projects D-Scribe and EGRAPSA use advanced machine and deep learning to extract palaeographical information from papyri and achieve automatic writer classification.[20] Also noteworthy is Papy-S-Net, a network developed in the framework of the GESHAEM Project, designed to automatically match papyrus fragments.[21]

The ultimate goal is to move towards a "low-to-no human intervention for annotating images."[22] The automatic extraction of data is currently quite advanced for modern documents. However, additional challenges arise for the analysis of historical documents due to several factors, such as their frequent bad state of preservation, the intrinsic characteristics of the mediums, and, in the case of handwritten documents, the irregularity of the writing.[23] As a result, efforts towards implementing a fully automated detection and recognition of visual and textual features of historical documents are still ongoing.

While we wait for the machines to learn their fair share of data and be trained to process and provide accurate and usable data, it is possible to rely on simpler technologies and a fair-share of human labor to aid researchers in the collection of data on the material and visual features of papyri. The "Measurement Tool" was developed to provide

17 Preliminary work on the present project were presented by Gabriel Nocchi Macedo at the 30th Congress of Papyrology held in Paris (25–30 July 2022).

18 Document Image Analysis (DIA) is a field of study that focuses on the automatic processing and interpretation of documents starting from their image form. The goal of DIA is to extract meaningful information, such as text, graphics, tables, and structure, from these images. DIA combines techniques from image processing, pattern recognition, machine learning, and computer vision to achieve its objectives.

19 A systematic literature review of image datasets for document image analysis, focusing on historical documents, such as handwritten manuscripts and early prints is offered in Nikolaidou – Seuret – Mokayed – Liwicki 2022, 305–38.

20 An overview of the project is presented in Marthot-Santaniello 2021, 2; for the latest advancement and achievement of the project see Marthot-Santaniello – Tu Vu – Serbaeva – Beurton-Aimar 2023; Seuret – Marthot-Santaniello – White *et al.* 2023; Cilia – D'Alessandro – De Stefano *et al*. 2024, 422–36.

21 Papy-S-Net, see Pirrone – Beurton-Aimar – Journet 2019 and 2021.

22 Pirrone – Beurton-Aimar – Journet 2021, 219.

23 Nikolaidou – Seuret – Mokayed - Liwicki 2022, 305–6; Christlein – Marthot-Santaniello – Mayr *et al.* 2022.

a practical solution for the systematic collection of basic material, visual, and graphic aspects of ancient documents, with the scope of meeting the current research needs.[24]

2 The Measurement Tool

The technology behind the Measurement Tool allows to measures the size of these features by employing a bounding-box technology.[25] The system of annotation is semi-automated, as it requires the user to manually draw a box around the feature that needs to be measured. The measurement is carried out based on the number of pixels of the digital image, starting from a known unit of measure. It is necessary to calibrate the unit of measure using a reference object of known size within the image. Digitized images of papyri are often provided with a ruler. By measuring the reference object (in this case, the ruler), the program calculates the scale factor, which is the physical size of the reference object divided by the number of pixels it occupies (e.g., 5 cm / 250 pixels = 0.02 cm per pixel). Next, in order to find the physical size of a document, the program multiplies its pixel count by the scale factor (e.g., if the object is 100 pixels wide, then its physical size is 100 pixels x 0.02 cm per pixel = 2 cm). This method allows for accurate size determination of objects in images, provided the images are taken with minimal perspective distortion and the reference object is measured with a good degree of accuracy.[26] Measurements can be performed on written documents for which a digitized image is available online.

At its present stage of development, the tool can be used to measure basic material features of basic documents: Height, Width, Margins, Line Height, Interlinear Space, *Kolleseis* and *Kollemata*. However, it has limitations when dealing with documents that

24 The Measurement Tool was developed as part of my doctoral research on the materiality of administrative documents from Roman and Byzantine Oxyrhynchus, in the framework of the EVWRIT project, a research initiative led by Prof. Klaas Bentein and funded by the European Research Council (ERC). I extend my gratitude to Klaas Bentein, for encouraging and following the conceptualization of the present tool at all the steps of its realization. I also thank the Center for Digital Humanities of Ghent University, in particular Pieterjan Potter, and the Trismegistos Project, in particular Tom Gheldof, for developing the tool and addressing the issues and numerous questions that arose during its testing and use.

25 Bounding-box technology is a method used in computer vision and image processing to identify and localize objects within an image or video. A bounding box is a rectangular box that can be drawn around the object of interest, defining the region where the object is located. This technology is widely used in various applications, including object detection, recognition, tracking, and image segmentation. Modern object detection algorithms, such as YOLO (You Only Look Once), Faster R-CNN (Region-based Convolutional Neural Networks), and SSD (Single Shot MultiBox Detector), can automatically detect and draw bounding boxes around objects in real-time without human intervention: these programs adopt a fully-automated bounding-box technology. Among the most recent application of bounding-box technology to the study of papyri, see Cilia – De Stefano – Fontanella *et al.* 2021.

26 See n. 2.

present more complex structure (e.g. multiple textual areas or writing in the margins) or composite rolls (i.e. *tomoi synkollesimoi*). I will address these limitations in more detail later.

As mentioned above, recent guidelines for editing papyri emphasize the importance of providing a detailed physical description of the document. This description should include information about both the writing support (e.g., material and format of the papyrus) and the layout of the text (e.g. margins, visual structure). However, most published papyrological editions include this information quite sporadically. The lack of consistent physical descriptions in these editions has created a significant gap for research on topics related to materiality. The Measurement Tool aims to help scholars in the systematic collection of this valuable information and contribute to expanding the availability of larger sets of data for future researches on the materiality of papyri. In the following sections, I will provide a brief description of the features that can be measured using the Measurement Tool and will briefly emphasize the importance of each feature.

2.1 Height and Width

Information regarding the size of the height and width of a document are consistently provided in the context of papyrological edition.[27] Only a few among the earliest papyrological volumes lack information on the size of the documents.[28] Notwithstanding the availability of information regarding the height and width of documents in the printed edition, the systematic analysis of the materiality of papyri has shown that a function for the measurement of the documents might prove necessary for a number of ancillary activities related to their size. First, the tool allows to obtain the size of a document in the rare occasion this information is missing in the edition or when a new document must be measured – provided that a digital image is available. Second, the tool allows to verify the information offered in the edition, when the measurements are suspected to be wrong: one conspicuous case is P.Oxy. XLII 3047, a declaration of *abrochia*, the height of which is wrongly reported at 40.5 cm, instead of its actual 31.5 cm. More importantly, a tool to measure the size of documents proves to be particularly helpful whenever it is necessary to supplement the information offered in the edition. I refer in particular to the case of *tomoi synkollesimoi*. Editions customarily report the overall measure of the composite roll, but do not offer information over the size of the single documents. Let us consider P.Oxy. LXXIV 4996–4998 (Fig. 1), a *tomos* with four

27 One problematic aspect is that the measurements of the two dimensions in some editions of papyri are recorded without specifying which is the height and which is the width. Modern editorial guidelines attempted to streamline this practice, suggesting the use of the abbreviation W. (for Width) and H. (for Height) next to the corresponding value. See "Guidelines for editing papyri."

28 See e.g. P.Lond. vols. II–IV.

notifications of death: three of the four documents were edited – the first item in the *tomos* being too fragmentary – and a different publication number was attributed to each of them. Despite the independent edition of each item in the *tomos*, the editor only recorded the overall size of the four items glued together, 19 x 33 cm.[29] This corresponds respectively to the total width of the *tomos* and the height of the tallest document in the *tomos*, despite that fact that the introduction states that "the *kollemata* have different heights; the bottom parts are aligned, but the tops vary visibly (the fourth *kollema* is approximately 4 cm shorter than the second)." The measurement tool allows to readily measure the size of each single items in the *tomos*,[30] providing useful information over the original written artefacts before their processing.

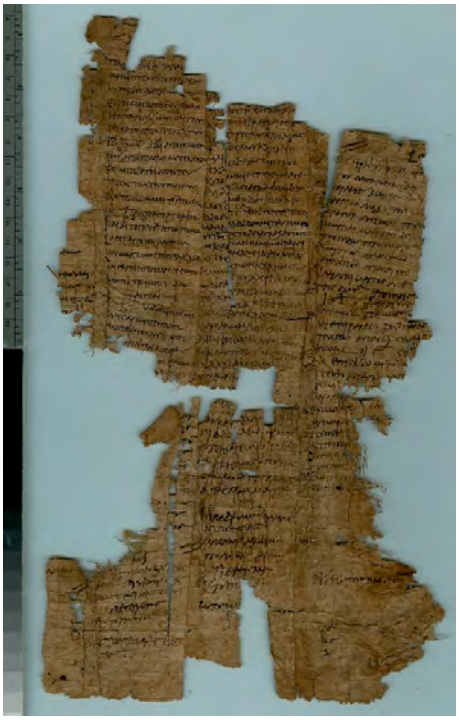


Fig. 1: P.Oxy. LXXIV 4996–4998.

29 P.Oxy. LXXIV 4996–4998, Introd., p. 119.

30 Namely 4996: h. 33 x w. 5.3 cm; 4997: h. [30] x w. 5.4 cm; 4998: h. 28.5 x w. 6.5 cm. To the width of each of these items it is necessary to add approximately 1 or 2 cm for the size left margin, which is obliterated by gluing the preceding document onto it, as customary for the creation of *tomoi synkollesimoi*. The lack of a digital image of the back of the *tomos* does not allow to calculate the exact size of the overlapping areas.

2.2 Margins

The distance between the text area and the edges of the sheet constitutes the margins of the document, which define its layout. The measures of the margins are found sporadically in older editions of papyri, and more often in recent ones.

Several elements collectively define the layout of a document, but the margins are the most prominent features that contribute to shaping the design of a text. In particular, margins define the proportion between negative (blank) and positive (written) space. This proportion is seen to vary considerably according to the formality of the document.[31] Margins are employed to create space around the text, making information easier to access; however, they are also functional areas: the blank space that they define around the text is often used for additional notes in private communication[32] or is often destined to receive additional official noted in administrative documents.[33] Their presence – or absence – can help us better understand the status of preservation of a document or its stage of production.

2.3 Line height

Information on the height of the lines in a document is found very sporadically in papyrological editions of documentary papyri, but it is more frequently found in editions of literary ones. The line height is strongly influenced by the writing style and expertise of the writer, but it also varies considerably during the centuries. Later in this contribution, I will present some preliminary results from the collection and analysis of data on line height and interlinear space using the Measurement Tool.

2.4 Interlinear Space

Just like the height of the writing line, the height of the interlinear space is also very rarely recorded in editions of documentary papyri, and more often in editions of literary ones. The height of the interlinear space is crucial in shaping the layout of the document and it varies considerably over the centuries and according to the writing style. Later in this contribution, I will present some preliminary results from the collection and analysis of data on line height and interlinear space using the Measurement Tool.

31 See e.g. the famous order from the prefect Aquila, SB I 4639.

32 Sarri 2018, 112–3.

33 I discuss the use and evolution of margins in petitions and declarations in my doctoral dissertation.

2.5 Position of the *kolleseis*

The presence and position of *kolleseis* (the joints between individual sheets of papyrus to form a roll) are only occasionally recorded in papyrological editions. This aspect is greatly understudied and certainly deserves more attention. While it is often said that *kolleseis* are randomly located on the documents,[34] there are no extensive studies that can support or deny this assumption. In documentary papyri, I have observed a certain tendency to find the *kollesis* rather close to the edges of the sheet and, in some noticeable cases, the scribe seems to be quite mindful of the *kolleseis*, using them to define the space of the writing column. One conspicuous example is P.Oxy. XXXIII 2666 (AD 308/9), where the two *kolleseis* seem to be strategically located in correspondence with the beginning and the end of the two writing columns (Fig. 2)[35] The subject shows great potential to uncover new aspects of writing practices on papyrus.

Fig. 2: P.Oxy. XXXIII 2666

34 Johnson 2004, 88.

35 See also e.g. PSI XII 1235 (AD 86), PSI IV 281 (1st cent. AD).

2.6 Width of the *kollemata*

The identification of the position of the *kolleseis* allows to measure the width of the *kollemata* (the sheets of papyrus joined to manufacture a roll). The Measurement Tool allows to record the width of one or more *kollemata* within a document. This information is only occasionally recorded in papyrological editions. According to Pliny, this feature is an important indicator of the quality of the papyrus roll.[36] To date, only partial and sparse investigation on the subject have appeared.[37] A systematic study of this aspect could significantly enhance our understanding of the manufacture of the rolls and offer valuable insights into the material culture, including the availability and selection of different qualities of writing supports in the Egyptian writing milieu.[38]

3 User Interface

The measurement tool is divided in two main sections:

– General Section: for the measurements of general features regarding the physical aspects of the document, the layout and the typography (size, margins, line height and interlinear space);
– *Kollemata* Section: for the measurement of the *kollemata* (position of the *kollesis*, width of the *kollemata*).

In both sections, the user interface consists of three main areas, as shown in Figure 3:

[1] Image Box: for the visualization of the digital image. It is possible to take some basic actions on the image using the buttons located in the top-left corner inside the box, such as zooming in or out (+ and –), bringing the image back to its original frame (home button) and expanding the area of the Image Box (Full Screen button). These actions will help the user to better navigate the image for a more efficient measurement of the document.

36 Plin. *NH* XIII 71–80. According to Johnson 1993, 48 this may be related to the fact that larger *kollemata* means less joints – *kolleseis* – in the writing area, therefore less imperfections to the surface and less impediment for the writer.

37 For a collection of small sets of data on the width of *kollemata* see Turner 1977, 48 in the study of the early codex; Turner 1978, 61; Robinson 1979, 9–46 in the study of the Nag Hammadi codices; Johnson 2004 in the study of literary rolls. I have conducted a preliminary study on this feature based on the administrative material from Oxyrhynchus in the framework of my doctoral dissertation.

38 Preliminary investigations of the administrative material from Oxyrhynchus, compared with the data provided by Johnson 2004 for literary rolls, suggest a deliberate choice of writing support by the writer according to the genre.

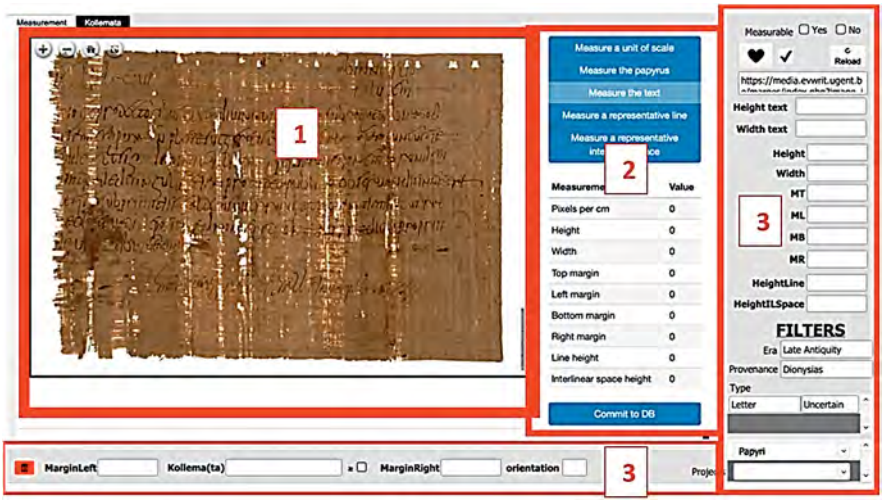


Fig. 3: Measurement Tool's interface.

[2] Command Box, which includes the command buttons to perform the measurements; The commands differ between the "General" section and the "*Kollemata*" section (Fig. 4).

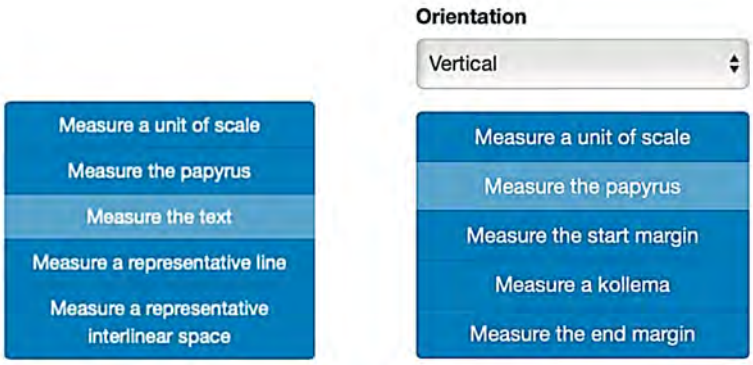


Fig. 4: Command Box, "General" section (left) and "*Kollemata*" section (right).

[3] Database Box: where the obtained measurements are displayed and become available for future searches. This box includes (1) general information on the document (Date, Provenance, Type, Height and Width) as recorded in the edition;[39] (2) information

39 These pieces of information are linked to the annotation available in the EVWRIT Database.

acquired through the Measurement Tool (Height, Width, MT, ML, MB, MR, HeightLine, HeightILSpace, MarginLeft, *Kollema*(ta), MarginRight, Orientation). It is possible to perform filtered searches by each of these fields.

4 User Instructions

A necessary requirement before performing any measurement is the calculation of a unit of measure, using the button "Measure a unit of scale" in the Command Box. Therefore, it is important that the digital image of the document is provided with a ruler. Digitized papyri are often provided with a ruler next to the image. The user must select the unit of measure on the ruler (Fig. 5) and proceed to specify the orientation of the ruler (whether it is placed horizontally or vertically); the number of units selected (one or more) and the metric system of the ruler (centimeters or inches). On the basis of this information the program will gauge the number of pixels per cm.

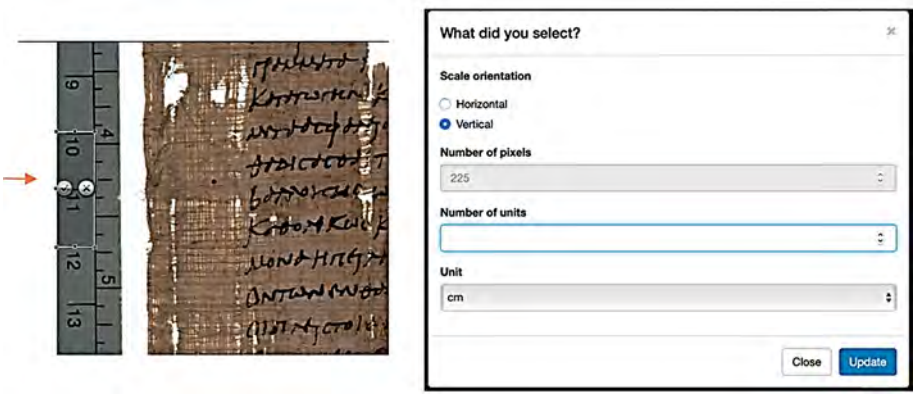


Fig. 5: Selecting a unit of measure.

4.1 General Measurement Section

4.1.1 Measure the Papyrus

This function allows the user to measure the Height and Width of the document. The measurement is performed using the button "Measure a papyrus" in the Command Box and then selecting the perimeter of the papyrus in correspondence to the widest and

the tallest points[40] (Fig. 6). The user can use the + and – buttons in the Image Box to enlarge and to move across the image in every direction, in order to make the selection as precise as possible. If the document is slightly tilted, it is possible to adjust the inclination of the selection. It is not yet possible to measure documents constituted by discontinuous fragments (see §6).

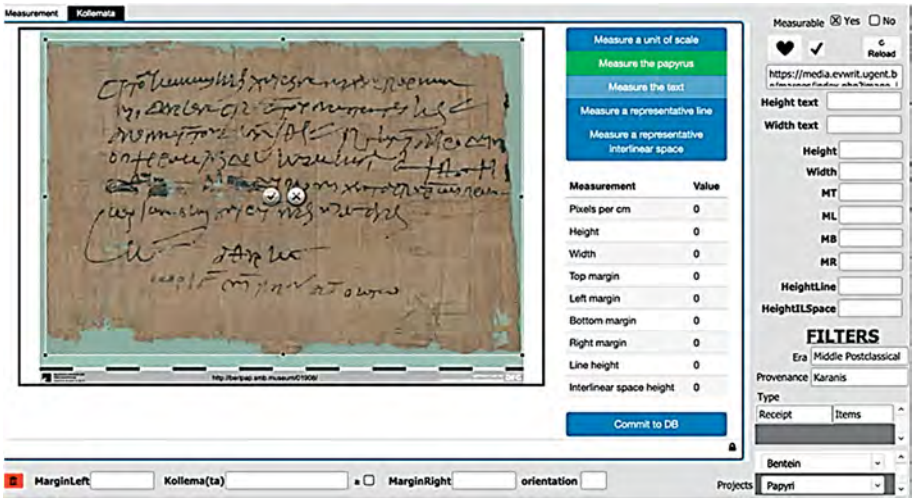


Fig.6: Measuring the papyrus.

4.1.2 Measure the Text

Measuring the text area is functional to obtaining the measurement of the Top, Bottom, Left and Right Margins of the document. For this reason, before proceeding to measure the text, it is essential to have previously measured the overall perimeter of the papyrus. The text area is measured by using the button "Measure the text" in the Command Box and drawing a box around the perimeter of the text (Fig. 7). If the perimeter of the text area is slanting, it is possible to adjust the inclination of the selection.

40 The same procedure applies to fragmentary documents that may present an irregular shape and to ostraca, see Fournet 2022, § 1.1.3: "When the document is incomplete, the dimensions are measured by inscribing the fragment (with the text lines in horizontal position) in a square or rectangle whose two horizontal lines correspond to the most extreme upper and lower points of the fragment and the two vertical lines to the most distant lateral points. In the case of an ostracon, whose original edges were rarely straight, it is conventional to consider as height the distance between the two upper and lower extremities of the sherd (when oriented so that the writing is horizontal) and as length/width the distance between the two lateral extremities."

When the area of the text has been measured, the program will automatically calculate the size of the margins. Currently, the Measurement Tool can only record the value of one textual area per image. Therefore, it is not possible to measure e.g. the area of multiple columns, independent texts on the same sheet or writing in the margins (see §6).

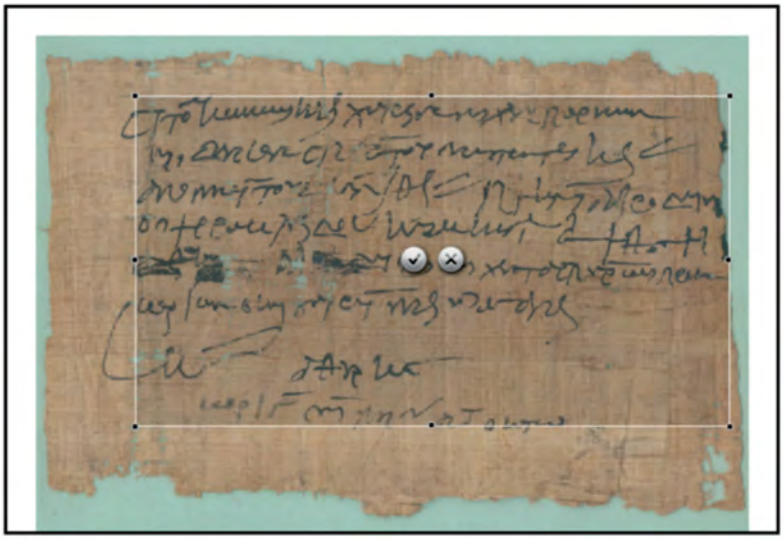


Fig. 7: Measuring the text area.

4.1.3 Measure a representative Line/Interlinear Space

This function allows to measure the height of the line or interlinear space, by selecting the corresponding portion of a line or interlinear space (Fig. 8). If the writing line is slanting, it is possible to adjust the inclination of the selection.

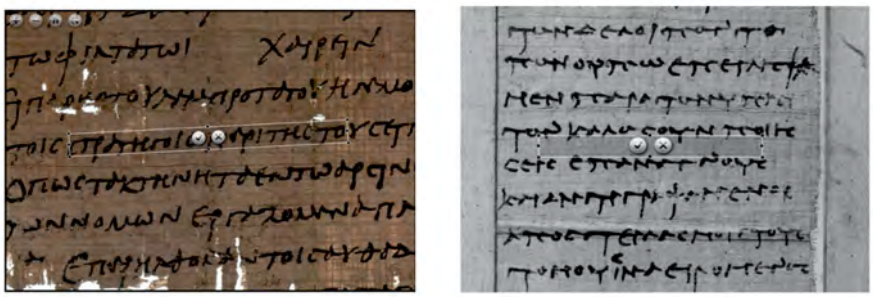


Fig. 8: Measuring a representative line (left) or Interlinear Space (right).

A certain degree of variation in the line height or the interlinear space must be expected even in the most regular handwriting. However, it is currently impossible to measure the average line height or interlinear space based on multiple lines within the document nor to record the height of different lines or interlinear space, when their size differs within the same document[41] (see §6).

The criteria for the selection of the height of these typographical features call for an in-depth discussion which goes beyond the scope of the present analysis. The main challenge arises from the wide range of graphic styles, which make it difficult to determine whether in measuring the distance between the writing lines (interlinear space) one should take the body of the letters as start and end point or include their vertical extension. Similarly, it is hard to evaluate whether the height of a writing line should be defined solely by the body of the letters (bilinear system) or whether it should also account for the vertically elongated traits (quadrilinear system). In the preliminary analysis presented later, I have chosen to use the bilinear system to calculate the height of the writing line and interlinear space.[42]

4.2 "*Kollemata*" Measurement Section

In this section of the Measurement Tool, it is possible to calculate and record the position of the *kolleseis* in the document and the width of *kollema*(*ta*), if any.

In order to start the measurement, the user first has to select the orientation of the *kollesis* (vertical or horizontal) in the Commands Area (Fig. 9). This may vary according to the orientation of the writing support (e.g. if the document is written transversa charta the *kollesis* will be horizontal) and the side of the roll on which the text is written (e.g. if the document is written transfibrally on the verso, the *kollesis* will be horizontal).

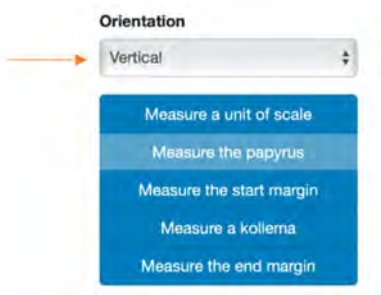


Fig. 9: Commands Area.

41 See e.g. P.Oxy. XXXIII 2673.

42 A different system is adopted in a cases of automatic layout analysis for the extraction of text height, presented in Pintus – Yang – Rushmeier 2013.

4.2.1 Measure the Start/End Margin

The position of the *kolleseis* on the sheet is usually calculated by measuring their distance from the nearest edge.[43] In the context of the Measurement Tool, the distance between the left edge and the nearest *kollesis* (if any) is defined as "Start Margin."[44] The distance between the right edge and the nearest *kollesis* (if any) is defined as "End Margin".[45] The distance between the two is the *kollema*. For example, P.Oxy. IX 1204 (Fig. 10) presents two *kolleseis*, one close to the left edge, which define the start margin, and one close to the right edge of the document, which defines the end margin.

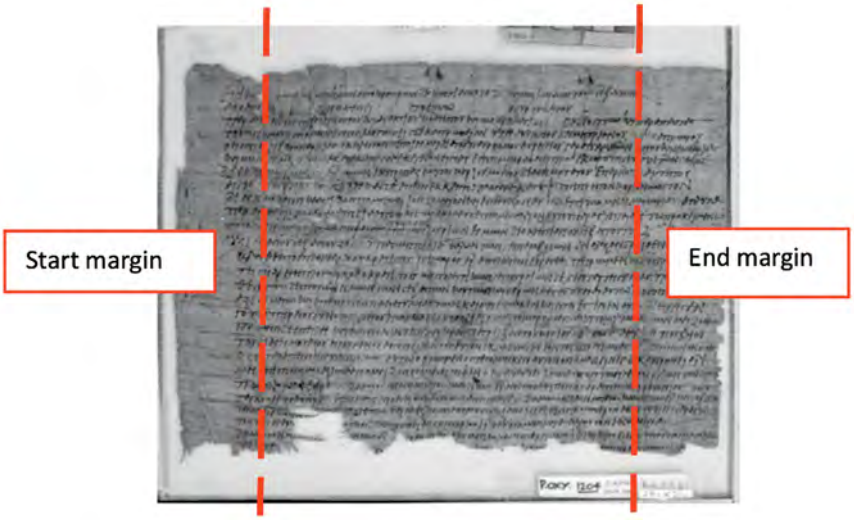


Fig. 10: P.Oxy. IX 1204.

In order to measure the Start Margin, the user will press the button "Measure Start Margin" in the Command Box. Then he will proceed to select the width of the area of interest: it is not necessary to select the entire height of the concerned area, as long as the selected width is accurate (Fig. 11). By measuring the Start Margin, the tool calculates the position of the *kollesis* in relation to the edge.

The same procedure can be repeated to calculate the End Margin, by pressing the button "Measure End Margin" and selecting the corresponding area.

43 See Fournet 2022, § 1.1.4.
44 To distinguish it from the Left Margin.
45 To distinguish it from the Right Margin

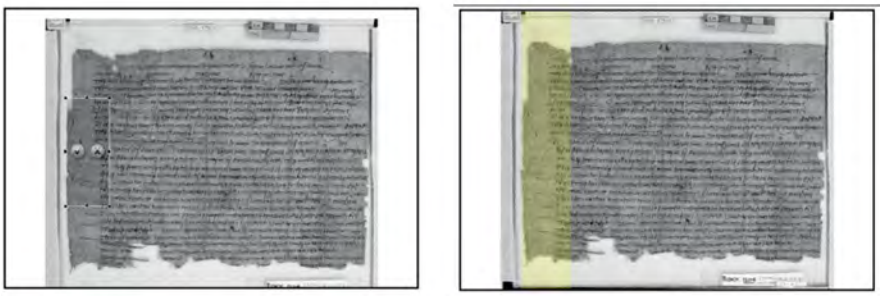


Fig. 11: Measuring “Start Margin”.

4.2.2 Measure a *Kollema*

The Tool allows to measure and record the width of one or more *kollemata*. It is also possible to perform the measurement when the *kollema* is only partially preserved.

(1) *Kollema* entirely preserved: when it is included between two *kolleseis* (Fig. 12).

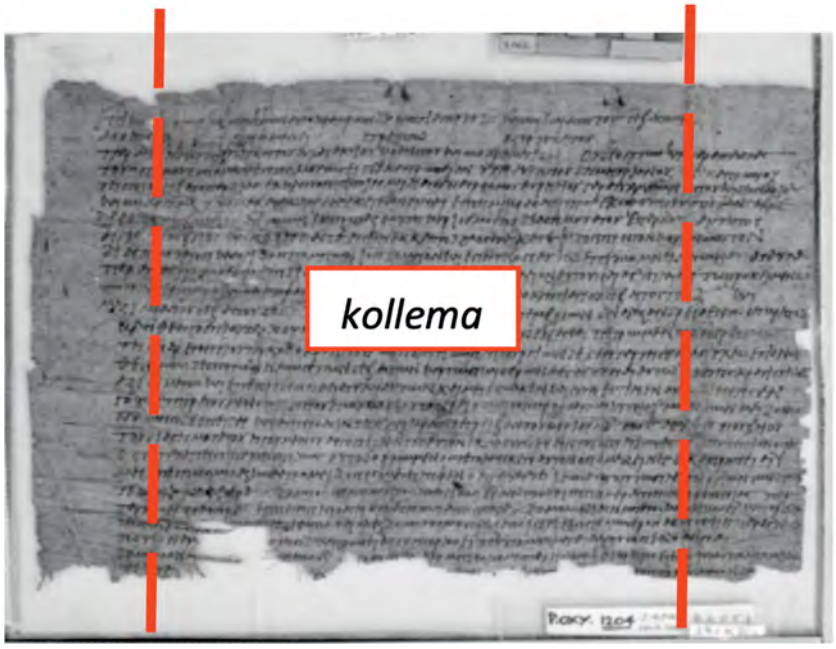


Fig. 12: *Kollema* entirely preserved.

In order to proceed to the measurement, the user will press the button "Measure a *kollema*" in the Command Box and proceed to the selection of the corresponding area on the image. In order to measure the *kollema*: it is sufficient to select the exact width of the concerned area (Fig. 13).

Fig. 13: Measuring the width of a *kollema.*

The size of start margin, *kollema*(ta) and end margin appear in the "Database Box," in the bar positioned in bottom part of the Measurement Tool (Fig. 14).

Fig. 14: Database Box.

This display format follows the guidelines recommended for presenting information about *kolleseis* and *kollemata* in papyrus editing (e.g. 4/19.1/4.4 cm).[46] It reads: two *kolleseis*, one 4 cm from the left margin and one 4.5 cm from the right margin, with a resulting *kollema* of 19.1 cm.

(2) *Kollema* partially preserved: when the document presents only one *kollesis* (Fig. 15) or no *kolleseis* at all.[47] In both cases, the original width of the *kollema* is lost, but it is still possible to extract information on its minimum measurable width. This information can be of interest, in particular when the partially preserved *kollemata* are significantly wide despite being incomplete. One relevant example is P.Bingen 78, a portion of a roll containing legal proceedings: the document is broken on the left and right

46 See Fournet 2022; see also Fournet 2019. Such system of display was first suggested by Maehler in the introduction to P.Oxy. LVIII 3985.

47 This happens when the document was written on a portion of a roll too small to encompass two *kolleseis*, or when it suffered damages that caused the loss of part of the document.

side, measures 46.5 cm in width, and only one single *kollesis* is visible at approximately 23 cm from the left edge. This results in two partial *kollemata* of 23 and 23.5 cm, respectively.

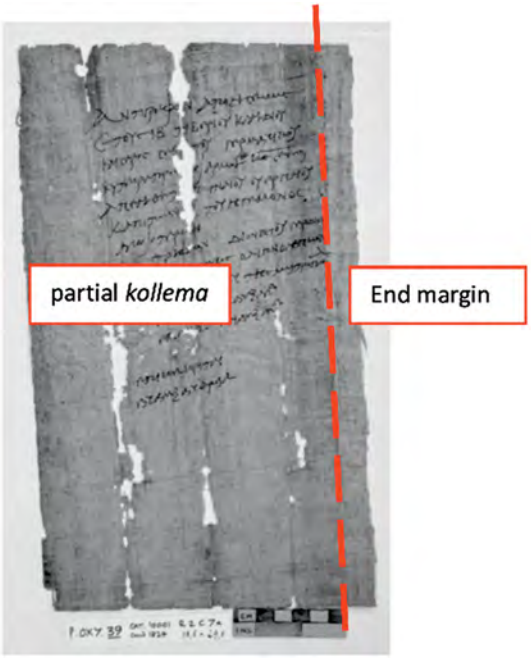


Fig. 15: *Kollesis* partially preserved.

If at least one *kollesis* is preserved, the user can proceed to measure the Start/End Margin[48] and the width of the partial *kollema*. It is possible to indicate that the value does not correspond to the entire width by ticking the ≥ box, next to the corresponding *Kollema*(ta) field. This indicates that the width of the *kollema* was originally greater or equal to the portion preserved. If no *kollesis* is present, the user can measure the width of the partial *kollema*, ticking the ≥ box, next to the *Kollema*(*ta*) field, to indicate that the width of the *kollema* was originally greater or equal to the portion preserved (Fig. 16).

Fig. 16: Database Box.

48 The missing margin will not be measured and will result in a "0" value.

5 Preliminary results: the evolution of height and interlinear space in petitions from Oxyrhynchus

As part of my dissertation, I have collected and conducted a preliminary analysis of each of the material features mentioned above. The full results of this work will be the object of future publications. In the meantime, to demonstrate the relevance of quantitative analysis in the study of the materiality of papyri and to showcase the potential of the Measurement Tool for the collection of data, I will briefly present some initial findings related to the height of the writing lines and the interlinear space here below.

Although these two typographical features are mainly regarded as stylistic features of a document, they are crucial aspects in the evolution of petitions. Their development is often interpreted from a palaeographic perspective, as a reflection of cultural trends in the writing style. However, their size also plays a central role in the materiality of the document and the evolution of format over time.

The dataset includes the line height and interlinear space of Oxyrhynchite petitions from the first until the 7th century AD, for a total of 196 petitions.[49] The documents are chronologically distributed as follows: 32 cases for the 1st century (16.3%); 28 cases for the 2nd century (14.3%); 59 cases for the 3rd century (30%); 39 cases for the 4th century (19.9%); 24 cases for the 5th century (12.2%); 13 cases for the 6th cen-tury (6.6%) and one single case for the 7th century (0.5%).

The height of the writing line is calculated by adopting the principle of bilinearity:[50] I considered only the body of the letters contained between two horizontal lines and disregarded the traits extending up or down, respectively above or below the horizontal lines in the interlinear space (Fig. 17). Similarly, the interlinear space is calculated by taking into account the distance between the body of the upper and bottom line, disregarding possible elongated traits of the letter extending into the space (Fig. 18). A bilinear system offers a greater consistency in the collection of data, especially in the context of cursive hands, which can present a high degree of irregularity in the orientation of the script.[51]

49 Out of a total of 302 petitions published to date from Oxyrhynchus for the considered period.

50 For a discussion on the formal features of the handwriting, see Amory forthcoming.

51 I adopt the definition of orientation offered by Amory forthcoming: "orientation denotes if the typefaces are elongated and therefore more vertically oriented, which can be perceived as 'light', or rather that they stretch horizontally, which can be seen as 'heavy.'

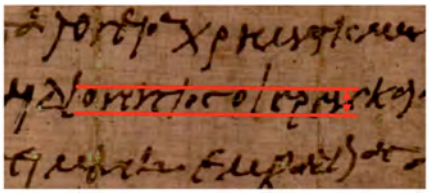


Fig. 17: Line height.

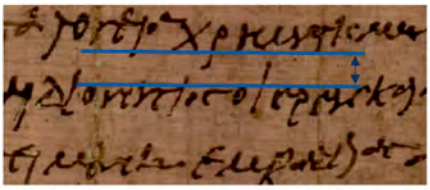


Fig. 18: Interlinear space.

The quantitative approach allowed to uncover patterns and trends not forcibly evident through the simple observation of the distribution of the text. The results are showcased in Figure 19, where the data on the line height (red dashes) is combined with that of the height of the interlinear space (blue crosses), to better illustrate the evolution of these features from the 1st century to the 7th century.

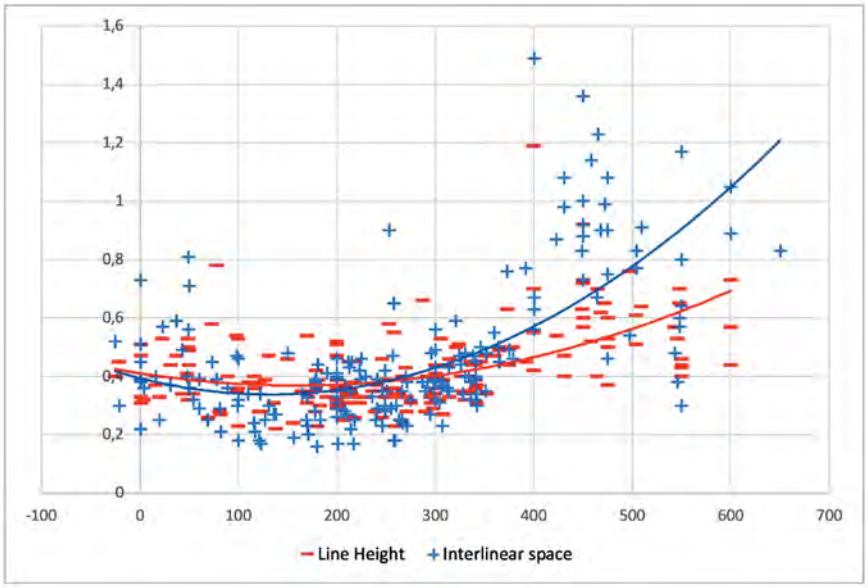


Fig. 19: The evolution of line height and interlinear space in petitions between the 1st and 7th century.

During the Roman period the line height remained rather stable throughout the centuries, ranging between ca. 0.20 and 0.60 cm, with an average between 0.35 and 0.40 cm.[52]

52 Line Height, Mean: 0.37; Median: 0.35; Mode: 0.33.

This trend is closely mirrored by the size of the interlinear space.[53] Interestingly, between the second and the middle of the 3rd century, a group of documents with a very reduced interlinear space below 0.20 cm appears in the chart.

In this period, it is possible to observe one main cluster, in line with the standard distribution of the text during the Roman period and characterized by lines and interlinear space of approximately 0.40 cm. This is well exemplified by P.Oxy. L 3555 (1st–2nd cent. AD), where both the line height and the interlinear space settle at around 0.39 cm (Fig. 20). Next to the main trend, a lower area appears in the Chart, characterized by documents with an extremely narrow interlinear space below 0.2 cm (Fig. 19). Developments and changes within this range might initially appear negligible, often being in the order of tenths of a centimeter. However, since handwriting and interlinear space are inherently small features, even minor variations can have significant implications.

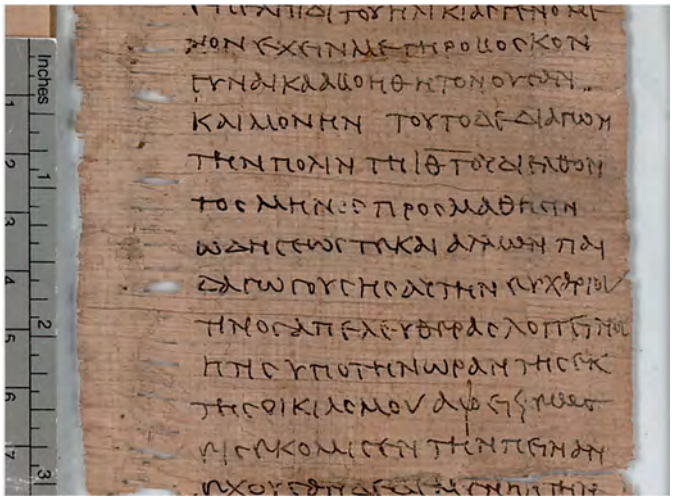


Fig. 20: P.Oxy. L 3555.

In fact, this is a period when the increasing complexity of administrative procedures, coupled with an intensification of the use of legal sources to support the claims, gave rise to the production of significantly longer petitions. To accommodate these extensive texts within the constraints of a single column, writers had to meticulously balance the proportions between writing line, interlinear spacing, and margins. As excessively reducing the height of the letters would compromise the document's readability, adjusting the interlinear spacing proved to be the most viable solution. As such, the interlinear

53 Interlinear Space, Mean: 0.35; Median: 0.35; Mode: 0.25.

space was often minimized to ensure that the text fit within its column.[54] A good example of this practice is P.Mich. XI 614 (AD 258/9), a long petition to recover a debt, with 0.30 cm line height and 0.18 cm interlinear space (Fig. 21).[55]

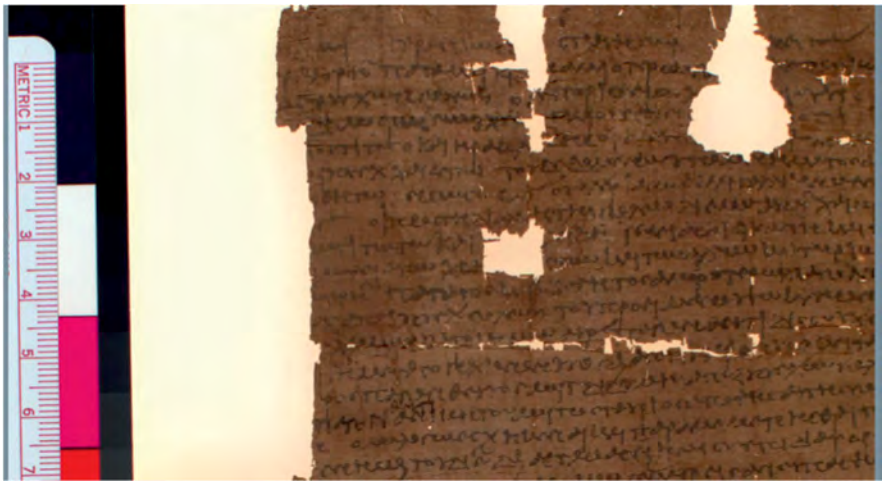


Fig. 21: P.Mich. XI 614

Around the end of the 3rd and the beginning of the 4th century, the sizes of both line height and interlinear space start to rise modestly, and eventually settle between 0.40 and 0.50 cm: these are prodromes of a change in writing style that will be fully implemented during the 5th century. This period corresponds to Diocletian's wave of reforms: the progressively wider use of Latin in the administrative production fostered the rapid influence of Latin on the Greek writing style and led to the development towards the so-called Byzantine cursive.[56]

In the 5th century the shift towards the sizeable Byzantine style handwriting is ultimately accomplished. This is evident from the size of the script: the line height has grown further – albeit only modestly – clustering between 0.40 and 0.70 cm.[57] More

54 See SB XVI 12698 (AD 180–192): 0.16 cm; P.Oxy. VI 898 (AD 123): 0.17 cm; PSI XIII 1328 (AD 201): 0.17 cm; P.Hamb. IV 271 (2nd cent. AD): 0.18 cm; P.Fouad 30 (AD 121): 0.18 cm; P.Mich. XI 614 (AD 258): 0.18 cm; SB XXIV 16265 (AD 259/60): 0.18 cm; P.Oxy. III 487 (AD 156): 0.19 cm; SB VIII 9905 (AD 171): 0.20 cm; P.Oxy. II 286 (AD 82): 0.21 cm; SB I 5678 (AD 117): 0.21 cm.

55 Similar for content, size of the document, handwriting and distribution of the writing lines is also PSI XIII 1328 (AD 201 C), which contained multiple attachments of previous phases of the procedure. The writer distributed the text in 75 lines of 0.30 cm in height but had to keep the interlinear space to the bare minimum, at merely 0.17 cm.

56 Cavallo 2008, 121 and bibliography mentioned there.

57 Mean: 0.55; Median: 0.55; Mode: 0.54.

conspicuous is the size of the interlinear space, which spikes up to reach heights between 0.70 and 1.2 cm. A good example is P.Köln. V 234 (AD 431), where the line-height measures 0.47 and interlinear space reaches 1.08 (Fig. 22).[58] This period is characterized in particular by a distinct increase of the interlinear space in comparison to line height.

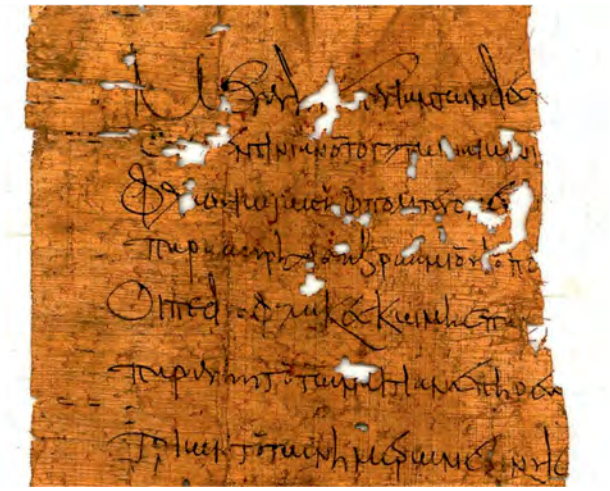


Fig. 22: P.Köln. V 234.

The evolution of the handwriting from the Roman to the Byzantine period is not simply a matter of size of the handwriting, but also – and perhaps more remarkably – a matter of proportions between the positive (written) and the negative (blank) space on the page. In order to show how this aspect of the document layout and typography changes, it is useful to calculate the ratio between the interlinear space and the line height, presented in Figure 23. As the Chart clearly shows, from the first until the 4th century this ratio oscillates between 0.5 and 1.5 – remaining closely around 1. In other words, the interlinear space and line height were equal or nearly equal in size.[59] Only from the 5th century onward, the ratio makes a conspicuous upward move, which points toward a clear change in the proportions between these two features.

58 Mean: 0.77; Median: 0.78; Mode: 0.83.

59 From the 1st until the end of the 4th century, the ratio mostly ranges between 0.5 and 1.4. This oscillation in value of the ratio is trivial, since it translates a difference of 0.9.

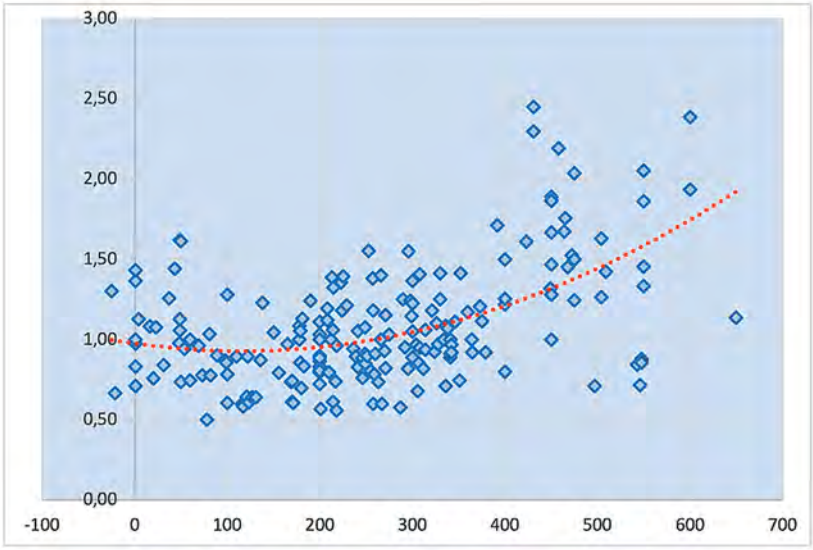


Fig. 23: Ratio evolution interlinear space: line height between the 1st and 7th century.

Analyzing the evolution of height and interlinear space can provide insights into the evolution of writing practices on a micro level. Typically, changes in writing practices are not marked by abrupt shifts but rather emerge as a series of small, incremental adjustments. While the evolution in the typography of documents may appear to be trivial or inconsequential, a quantitative analysis helps to highlight its significance and offers a solid ground for a deeper understanding of the material evolution of ancient documents.[60]

6 Limitations and Future Perspectives

At present, the Measurement tool is at its first stage of development. As it was designed to measure the essential features of documentary papyri with a basic format and layout, the range of material and visual variation found in the papyrological documentation and the complex structure of some documents will require additional work in order to further improve the present tool and adjust it for the measurement of more complex features. In particular, the Measurement Tool, is not equipped to measure:

(a) Documents constituted by multiple items: this is the case with documents with discontinuous fragments or *tomoi synkollesimoi*. Currently, the Measurement Tool can

60 I discuss the relationship between typography and format in my doctoral dissertation.

only record one set of measures (height, width, margins) per image. While it is possible to measure the individual fragments, it is not yet possible to store and retrieve the measurement of each distinct item as linked to one single image and TM number.

(b) Documents with multiple columns or complex layout: it is only possible to measure one text area per image. The tool is not equipped to measure and store values related to several columns or multiple textual areas (e.g. writing in the margins), if linked to one single TM number. Similarly, in the case of documents with multiple textual levels (e.g. documents which contain distinct documents on the same sheet/roll or documents with attachments) or of documents written on recto and verso (opistograph), it is currently not possible to measure and differentiate between the different parts.

(d) Average line height or multiple line heights/interlinear space: as discussed above, it is only possible to record the height of one representative line. It is currently impossible to record when the height of lines in the document changes throughout the text or if different written areas on the same sheet have different lines height.

(e) Text formatting: features such as *vacats*, line spacing, *ekthesis* or *eisthesis* cannot be measured.

The Measurement Tool is currently integrated into the EVWRIT Database, therefore it is only accessible to its users. Additional work and resources are needed to develop it as an independent, open-access resource and expand it functionalities. Given the current limitations of the Measurement Tool, future development will focus on expanding its capabilities to handle more complex documents. Additional functionalities could be implemented for the annotation of features related to text formatting (e.g. *vacats*, line spacing, *ekthesis* and *eisthesis*) and document folds.

More importantly, efforts will be directed towards making the Measurement Tool an independent, open-access resource, enabling wider accessibility and integration with other databases and analytical tools. This will facilitate more collaborative research and data sharing among scholars, ultimately advancing the study of the materiality of ancient documents.

7 Conclusions

In the analysis of the materiality of papyri, a full automatization of the process of collection of the data related to the physical, visual and typographical aspects of ancient documents is certainly desirable and perhaps in close sight. Among its many advantages, it would allow researchers to concentrate on the analysis and contextualization of the data, avoiding the tedious job of drawing boxes around the perimeter of thousands of documents, textual areas, writing lines or single letters, and it would likely offer cleaner data, free from the oversights and errors typical of manual annotations.

In the meantime, The Measurement Tool aims to address this drawback by providing a tool for a systematic collection of data on basic material and visual features of

papyri. While the tool currently has its limitations – in particular with the annotation of more complex material – its development marks a significant step forward in the study of materiality.

Future efforts should focus on developing the Measurement Tool into an independent, open-access resource, expanding its functionalities, and integrating it with other databases. This will enable more collaborative research and data sharing, ultimately enriching our understanding of ancient documents. By systematically addressing the challenges and harnessing the power of modern technology, we can look forward to a future where the study of ancient documents is more accessible, precise, and insightful than ever before.

Bibliography

Angliker, E. – Bultrighini, I. (2023), eds., *New Approaches to the Materiality of Text in the Ancient Mediterranean. From Monuments and Buildings to Small Portable Objects*, Turnhout.

Amory, Y. (forthcoming), *More than a Simple Intuition. Towards a Categorization of Paleographical Features in Greek Documentary Papyri*, in *Neo-Paleography: Analysing Ancient Handwritings in the Digital Age*, ed. by I. Marthot-Santaniello, to be published as a special issue of the Comparative Oriental Manuscript Studies Bulletin.

Bentein, K. – Amory, Y. (2022), *Introduction: Developing a Historical Social-Semiotic Approach to Communication Practices in Antiquity*, in *Novel Perspectives on Communication Practices in Antiquity: Towards a Historical Social-Semiotic Approach*, ed. by K. Bentein – Y. Amory, Leiden – Boston, 1–14.

Cavallo, G. (2008), *La scrittura greca e latina dei papiri. Una introduzione*, Pisa – Roma.

Cilia, N. D. – De Stefano, C. – Fontanella, F. – Marthot-Santaniello, I. – Scotto Di Freca, A. (2021), *PapyRow: A Dataset of Row Images from Ancient Greek Papyri for Writers Identification*, in *Pattern Recognition. ICPR International Workshops and Challenges*, ed. by A. Del Bimbo – R. Cucchiara – S. Sclaroff – G. M. Farinella – T. Mei – M. Bertini – H. J. Escalante – R. Vezzani, Cham, 223–34. [https://doi.org/10.1007/978-3-030-68787-8_16] (all hyperlinks accessed on 21.7.2024)

Cilia, N. D. – T. D'Alessandro – C. De Stefano – F. Fontanella – I. Marthot-Santaniello – M. Molinara – A. Scotto Di Freca (2024), *A Novel Writer Identification Approach for Greek Papyri Images*, in Image Analysis and Processing - ICIAP 2023 Workshops, ed. by G. L. Foresti – A. Fusiello – E. Hancock, Cham, 422–36. [https://doi.org/10.1007/978-3-031-51026-7_36]

Christlein, V. – I. Marthot-Santaniello – M. Mayr – A. Nicolaou – M. Seuret (2022), *Writer Retrieval and Writer Identification in Greek Papyri*, in *Intertwining Graphonomics with Human Movements. 20th International Conference of the International Graphonomics Society, IGS 2021, Las Palmas de Gran Canaria, Spain, June 7–9, 2022, Proceedings*, ed. by C. Carmona-Duarte – M. Diaz – M. A. Ferrer – A. Morales, Cham, 76–89. [https://doi.org/10.1007/978-3-031-19745-1_6]

Ferretti, L. – Fogarty, S. – Nury, E. – Schubert, P. (2020), *Cession of Cleruchic Land: From Procedure to Format*, ZPE 215, 201–10.

Fournet, J.-L. (2007), *Disposition et réalisation graphique des lettres et des pétitions protobyzantines : pour une paléographie « signifiante » des papyrus documentaires*, in *Proceedings of the 24th International Congress of Papyrology, Helsinki, 1–7 August, 2004*, ed. by J. Frösén – T. Purola – E. Salmenkivi, Helsinki, I, 353-367.

Fournet, J.-L. (2019), *Some Thoughts on the Papyrological Edition*, in *Proceedings of the 29th International Congress of Papyrology, Lecce, 28th July–3rd August 2019*, ed. by M. Capasso – P. Davoli – N. Pellé, Lecce, 460–70. [https://hal.science/hal-03876774]

Fournet, J.-L. (2022), *Guidelines for Editing Papyri*, CE 97, 306–46.
Ghigo, T. – Nodar Dominguez, A. (2023), *Reading the Materiality of the Oxyrhynchus Papyri: Non-Invasive Analyses to Reveal Scribal Choices*, Archaeological and Anthropological Sciences 15, #132, https://doi.org/10.1007/s12520-023-01839-9.
Hoogendijk, F. A. J. – van Gompel, S. M. T. (2018), eds., *The Materiality of Texts from Ancient Egypt: New Approaches to the Study of Textual Material from the Early Pharaonic to the Late Antique Period*, Leiden.
Johnson, W. A. (1993), *Pliny the Elder and Standardized Roll Heights in the Manufacture of Papyrus*, Classical Philology 88, 46–50.
Johnson, W. A. (2004), *Bookrolls and Scribes in Oxyrhynchus*, Toronto.
Krutzsch, M. (2012), *Das Papyrusmaterial im Wandel der antiken Welt*, APF 58, 101–8.
Krutzsch, M. (2020), *Material-Technical Details on Papyrus as Writing Support*, in *Proceedings of the Third International Conference on Natural Science and Technology in Manuscript Analysis and the Workshop OpenX for Interdisciplinary Computational Manuscript Research. University of Hanburg, Centre for the Study of Manuscript Cultures (12–14 June 2018)*, ed. by O. Hahn – V. Märgner – I. Rabin – H. S. Stiehl, Hamburg, 123–32.
Marthot-Santaniello, I. (2021), *D-scribes Project and Beyond: Building a Virtual Research Environment for the Digital Palaeography of Ancient Greek and Coptic Papyri*, Classics@ 18, https://classics-at.chs.harvard.edu/classics18-marthot-santaniello.
Marthot-Santaniello, I. – M. Tu Vu – O. Serbaeva – M. Beurton-Aimar (2023), *Stylistic Similarities in Greek Papyri Based on Letter Shapes: A Deep Learning Approach*, in *Document Analysis and Recognition – ICDAR 2023 Workshops*, ed. by M. Coustaty – A. Fornés, Cham, https://doi.org/10.1007/978-3-031-41498-5_22.
Mazza, R. (2021), *Descriptions and the Materiality of Texts*, Qualitative Research 21, 376–93, https://doi.org/10.1177/1468794121992736.
Nikolaidou, K. – Seuret, M. – Mokayed, H. – Liwicki, M. (2022), *A Survey of Historical Document Image Datasets*, International Journal on Document Analysis and Recognition 25, 305–38, https://doi.org/10.1007/s10032-022-00405-8.
Petrovic, A. – Petrovic, I. – Thomas, E. (2019), eds., *The Materiality of Text – Placement, Perception, and Presence of Inscribed Texts in Classical Antiquity*, Leiden – Boston.
Pintus, R. – Yang, Y. – Rushmeier, H. (2013), *ATHENA: Automatic Text Height Extraction for the Analysis of Text Lines in Old Handwritten Manuscripts*, Journal of Computing and Cultural Heritage 8, 1–25, https://doi.org/10.1109/DigitalHeritage.2013.6743802.
Pirrone, A. – Beurton-Aimar, M. – Journet, N. (2019), *Papy-S-Net : A Siamese Network to Match Papyrus Fragments*, in *HIP '19: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, September 20–21, 2019, Sydney, Australia*, New York, 78–83, https://doi.org/10.1145/3352631.3352646.
Pirrone, A. – Beurton-Aimar, M. – Journet, N. (2021), *Self-Supervised Deep Metric Learning for Ancient Papyrus Fragments Retrieval*, International Journal on Document Analysis and Recognition 24, 219–34, https://link.springer.com/article/10.1007/s10032-021-00369-1.
Reggiani, N. (2024), ed., *Materialità della medicina antica. Aspetti grafici e materiali dei papiri medici dall'Antico Egitto*, Parma.
Rekrut, A. (2014), *Matters of Substance: Materiality and Meaning in Historical Records and Their Digital Images*, Archives and Manuscripts 42, 238–47.
Robinson, J. M., (1979), *Codicological Analysis of Nag Hammadi Codices V and VI and Papyrus Berolinensis 8502*, in *Nag Hammadi Codices V, 2-5 and VI: With Papyrus Berolinensis 8502, 1 and 4*, ed. by D. M. Parrott, Leiden, 9–46.
Sarri, A. (2018), *Material Aspects of Letter Writing in the Graeco-Roman World: c. 500 BC – c. AD 300*, Berlin – Boston.
Schubert, P. (2016), *On the Form and Content of the Certificates of Pagan Sacrifice*, JRS 106, 172–98.
Schubert, P. (2018), *Warrants: Some Further Considerations on Their Typology*, BASP 55, 253–74.

Schubert, P. (2022), *The Format, Layout and Provenance of Documents Pertaining to Liturgy*, Pylon 1, https://doi.org/10.48631/pylon.2022.1.89327.

Seuret, M. – I. Marthot-Santaniello – S. A. White – O. Serbaeva Saraogi – S. Agolli – G. Carrière – D. Rodriguez-Salas – V. Christlein (2023), *ICDAR 2023 Competition on Detection and Recognition of Greek Letters on Papyri*, in *Document Analysis and Recognition - ICDAR 2023*, ed. by G. A. Fink – R. Jain – K. Kise – R. Zanibbi, Cham, 498–507. [https://doi.org/10.1007/978-3-031-41679-8_29]

Turner, E. G (1971), *Greek Manuscripts of the Ancient World*, Princeton.

Turner, E. G. (1978), *The Terms Recto and Verso: The Anatomy of the Papyrus Roll*, Bruxelles.

Elisa Nury

The *grammateus* Project: Innovation and Challenges while Reusing Papyrological Data

1 Introduction

Papyrology is a field of research that adopted digital tools and methods as early as the 1960s and has seen the development of numerous databases and electronic resources: text corpora, such as the *Duke Databank of Documentary Papyri* (DDbDP),[1] or metadata about the material documents, such as the *Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens* (HGV).[2] Under the umbrella of a series of projects named *Integrating Digital Papyrology* (IDP), these two resources, along with APIS[3] and later Trismegistos[4] were brought together in the platform Papyri.info.[5]

With Papyri.info, researchers have access to a very large amount of papyrological data through two combined components: the *Papyrological Navigator* (PN), which is the search interface, and the *Papyrological Editor* (PE), a collaborative environment where anyone can update existing records or create new ones. Each change is peer-reviewed by the board of editors, ensuring the high quality of the data.

However, benefitting from the continuous work from the community is not straightforward. Other projects can include a copy of the data at a fixed point in time before adding new layers of interpretations. While this approach allows for tailoring the data to the needs of a particular project, once the data is transformed it becomes difficult to update it with the latest developments from Papyri.info. A risk of obsolescence is therefore present.

This article was prepared within the frame of two research projects funded by the Swiss National Science Foundation (grants #182205 and #212424) and based at the University of Geneva. The first phase of the project, *grammateus: the architecture of Greek documentary papyri*, ran from 2019 to 2023 with a team composed of prof. Paul Schubert (PI), Susan Fogarty, Lavinia Ferretti and myself. The project is currently in its second phase (*Greek documentary papyri: from architecture to periodization*, 2023-2026), with Ruey-Lin Chang and Gianluca Bonagura joining the team. I warmly thank my colleagues for their helpful remarks and suggestions.

1 https://papyri.info/docs/ddbdp (Accessed September 5, 2023).
2 https://aquila.zaw.uni-heidelberg.de (Accessed September 5, 2023).
3 https://papyri.info/docs/apis (Accessed September 5, 2023).
4 https://www.trismegistos.org (Accessed September 5, 2023), see Depauw – Gheldof 2014.
5 https://papyri.info (Accessed September 5, 2023). On the IDP projects and the creation of Papyri.info, see also Bagnall 2010; Sosin 2010; Babeu 2011, 141–56, 217; Baumann – Bodard – Cayless *et al.* 2011; Baumann 2013; Reggiani 2017, 222–40.

https://doi.org/10.1515/9783111070162-016

How can we best take advantage of the work from the ongoing collaborative effort to maintain and increase the corpus of edited texts and their associated metadata? Although an increasing emphasis is placed on Open Research Data and the archiving of fixed research data for future reuse, it is also worth considering the case of live data. This article will discuss the approach of *grammateus*, a project proposing a new typology of Greek documentary papyri. We have endeavoured to reuse papyrological data dynamically, in order to present the most up-to-date transcription, date and place of origin for the papyri in our database, while adding a new layer of interpretation to the text by highlighting sections. Here we present the challenges of this innovative approach, and make suggestions on how the transcriptions could be improved to facilitate such data reuse.

2 *grammateus*

The *grammateus* project aims to provide a wide survey of the attested Greek documentary papyri types. Classification attempts for documentary papyri have already been proposed in the past. For instance, Orsolina Montevecchi listed thirteen categories of documents, divided into subcategories and sometimes further detailed.[6] She described more precisely two of the broad categories, documents sent by private parties to state officials (no 6) and documents established between two private parties (no 7).[7] However, many categories, among other Private Life (no. 11, including letters) and Administration (no 2, including registers), are left undescribed. On the other hand, some narrow types of documents have been described extensively, such as wet-nurse contracts.[8]

More recently, Trismegistos also built its own classification underlying TM Texts, which was initially not displayed online and described as "without doubt the least standardised field in the database".[9] Currently, there is a beta feature that gives the content type of each specific document, e.g. P.Mich. I 34 (TM 1934) has content "Declaration: notification", and a search field to look for them. Fifty of the most common types are listed in the information available for the "Type" search field.[10] However, a full list of content types is not visible in Trismegistos, and there is also no public description of what makes a particular type and the criteria followed to classify documents into types.[11]

HGV subjects are keywords describing the content of a text and are even less standardised than the types of Trismegistos: they include not only information about the type

6 Montevecchi 1988, 86–9.

7 Montevecchi 1988, 177–233.

8 Manca Masciadri – Montevecchi 1984.

9 Depauw – Gheldof 2014.

10 See https://www.trismegistos.org/tm/index.php (Accessed October 4, 2023).

11 We are grateful to our colleague Joanne Stolk who kindly shared her classification with us.

of document, but also the names of persons involved such as senders and addressees, sometimes also a summary with full sentences describing the content.[12] Despite the many variations, there is clearly an outline of classification into types, with subtypes included between parentheses. Letters are often categorised as "privat", "amtlich", "geschäftlich", and sometimes two of those together; it is also possible to find among others "christlich", "kaiserlich" or "Empfehlung" as a qualifier. Uncertainty is indicated with question marks, and there can be small differences in capitalisation or spelling (e.g. "amtlicher" instead of "amtlich"). It is also possible to find an even more precise classification into sub-subtypes indicated in square brackets, for instance "Vertrag (Pacht [Land, Brachland, Sumpfland], Misthapoche)" (P.Poethke 28 [TM 128339]). Imperfect as it is, the HGV subjects are still an incredibly valuable tool to find documents of a certain type. For instance, Antonia Sarri included all letters marked as "Brief" in HGV to create the corpus of letters to study in her monograph.[13]

Grammateus aspires to offer a typology both of a wide scope and with detailed descriptions for each document category. To create this typology, we have attempted to follow a holistic approach, looking at all aspects of the construction of a document from the perspective of the scribe. The typology is based on the structure of the text (formulas), on its layout (for instance remarks about the use of *eisthesis* or *ekthesis*) and on format aspects such as the orientation of the page, direction of the fibres, and dimensions of the papyrus.

The premise at the core of *grammateus* is that scribes worked from models to build a document for a particular purpose. We have identified four types for general purposes: *Epistolary Exchange*, a two-way communication between sender and addressee; *Transmission of Information*, a one-way communication channel; *Objective Statement*, for simple acknowledgement of an action, e.g. tax receipts; *Recording of Information* for documents that record information on the longer term, such as accounts.

To establish the subtypes and their variations, we have also relied on the previous typology attempts, notably the Trismegistos content and the HGV subjects. After examining a broad range of documents from a category, we use the three elements of structure, format, and layout to outline a typical model and we select representative papyri to illustrate that model. We focus on complete documents, and not fragments, to analyse the material properties of the papyrus. The documents were selected examining mainly the content of HGV subjects and titles, and the presence of certain words in the texts.

12 As an example, the subjects of BGU IV 1078 (TM 9455) are: "Brief (privat); Sarapion an Sarapias; er beschwert sich; daß er nicht über die Abreise von Freunden unterrichtet worden ist und überlegt; was er arbeiten kann" (https://papyri.info/ddbdp/bgu;4;1078, accessed September 5, 2023). The semicolon separation marks how the subjects were separated into different <term>s in the XML encoding. Subjects are found under the label "Inhalt" in the HGV database and they are mainly in German although other languages can be found, e.g. English, French or Greek.

13 Sarri 2018, 53.

For instance, we reviewed among others the papyri containing the verbs ἐδάνεισε, ἐδάνισε or ἐμίσθωσε to prepare our selection of Syngraphe documents.[14]

Rather than produce a physical book, it was decided that it would be more useful to develop an interactive digital platform, accessing the huge amount of data already available on Papyri.info. Thus, the *grammateus* interface provides in its main section an introduction to our work, the typology with descriptions and information on the classification process, as well as a comparison tool. It is accompanied by a database of representative papyri selected to illustrate the typology realised by building a dynamic interface between the two sites. Finally, a detailed bibliography completes the website.[15]

3 Accessing Papyri.info: a dynamic reuse of data

Grammateus is not the first project to reuse existing papyrological open data. *Trismegistos Words* reused the transcriptions of Greek papyri to annotate words with part-of-speech (POS), morphological analysis and lemmata.[16] This new Trismegistos database was created from the existing transcriptions available on Papyri.info as of September 2016. The results are visible in TM Texts, where users can hover over words to see more information. *Callimachus* is "an automated regest of published papyri and ostraca, i.e. a processed extract of the formal contents of the text in the papyri hosted at the Papyri.info site"[17]. The database contains three kinds of information pertaining to documentary and literary papyri: countable features of the text such as words, letters, gaps, etc.; the state of the text; other metadata such as place and date of origin. The contents of Papyri.info were parsed, counted and annotated. A parallel database of morphologically annotated papyri and ostraca, *Polyphemus*, is also available.[18] *Sematia* is a corpus of annotated papyri and a platform to create such annotations in the form of treebanks. *Sematia* converts EpiDoc XML texts to a format that can be morphologically and syntactically annotated with the tool *Arethusa*.[19]

Whether clearly stated, as for TM Words, or implied, the projects mentioned above work from a fixed version of the Papyri.info corpus – or in the case of *Sematia* a fixed version of the texts, as the annotation work progressed while the number of Greek papyri kept growing on Papyri.info. For *grammateus*, we have decided to take a different approach to Papyri.info that we tentatively call "dynamic access".

14 Cf. Ferretti – Fogarty – Nury – Schubert 2023a. See Bonagura – Chang – Ferretti *et al*. 2023a; Ferretti – Fogarty – Nury – Schubert 2023b for more on the *grammateus* methodology and corpus selection.

15 https://grammateus.unige.ch (accessed September 13, 2023).

16 Keersmaekers – Depauw – Broux 2016.

17 https://web.archive.org/web/20220922203027/https://glg.csic.es/Callimachus/Callimachus_presentation.html (accessed September 13, 2023). See also D. Riaño Rufilanchas, this volume.

18 Riaño Rufilanchas 2023.

19 Vierros – Henriksson 2017; Vierros 2018.

Papyri.info offers an API to access the integrated papyrological data from HGV and DDbDP in XML EpiDoc standard in the form of URLs combined with identifiers.[20] For instance, the XML transcription of the papyrus known as P.Fay. 110 can be accessed with: https://papyri.info/ddbdp/p.fay;;110/source. The format to access a transcription is therefore to combine "http://papyri.info/ddbdp/" with the ddb-hybrid identifier and the string "/source". The ddb-hybrid identifier is made of the publication title components, the series name according to the checklist abbreviation (here P.Fay.), series volume and number separated by semicolons. On the other hand, the HGV record for any papyrus can be obtained by concatenating "https://papyri.info/hgv/" with the HGV identifier and "/source".

This API opens the way for accessing dynamically the XML data from Papyri.info through a list of proper identifiers. The interaction between *grammateus* and Papyri.info takes place on the pages displaying the papyri in our database: each time a page is opened, the XML is fetched from Papyri.info in order to display it on our interface, with different levels of transformation. The next section describes this process in more detail.

3.1 Content of a page

A papyrus page on *grammateus* displays first a set of metadata divided into three sections for identification information, format information, and links to other databases (Papyri.info, HGV, Trismegistos) and to the EpiDoc XML that we have produced. To facilitate access and reuse of our own data, we use the same stable URL schema, combining the *grammateus* URL with a TM identifier and "/source".

Below the metadata, a representation of the papyrus can be found, and when possible, a viewer with IIIF images to compare the model with the actual document.[21] The papyrus model is composed of three layers. First, a striped rectangle represents the papyrus proportionally to its original dimensions. The stripes indicate the direction of the fibres, horizontal or vertical. In the case of papyri that mix vertical and horizontal fibres, one half of the rectangle has vertical fibres and the other half horizontal ones. The mixed fibres visualisation applies to a very small subset of the database. It is not an accurate imitation of how the vertical and horizontal fibres intersect on the original papyrus, but only

20 The API is described by Cayless 2019, 39.

21 "IIIF is a set of open standards for delivering high-quality, attributed digital objects online at scale" (https://iiif.io, accessed September 15, 2023). At the start of *grammateus*, only the library of Manchester was offering images with IIIF manifests for papyri in our database. An increasing number of libraries are adopting this new technology, including the British Library and the libraries of Michigan, Hamburg, Bremen, Ghent, or Harvard universities, to name only a few. We currently show 544 images of papyri through IIIF. In addition, we also display 138 Oxyrhynchus papyri archived on Figshare: https://portal.sds.ox.ac.uk/Oxyrhynchus_Papyri (accessed September 11, 2023).

an indication that this phenomenon can be observed. The second layer is the text itself. Finally, the third layer is the colour highlighting of each text section.

Text sections are studied in *grammateus* as a part of the text's structure (see p. 263 above). Twelve sections have been identified as relevant to the typology.[22] Opening, Main Text, Closing and Date are general sections that can be found in all categories of documents, although the only one that is present in every papyrus of our database is Main Text. Other sections such as Column Number or Transitional Clause are related more closely to a certain type, in this case Lists or Registers. The presence of these sections and how they are organised on the page are of particular interest to us, which is why we highlight them.

Fig. 1 gives an example: upon opening the page, three elements are retrieved through Papyri.info: the provenance and century metadata from HGV, and the text transcription from DDbDP.

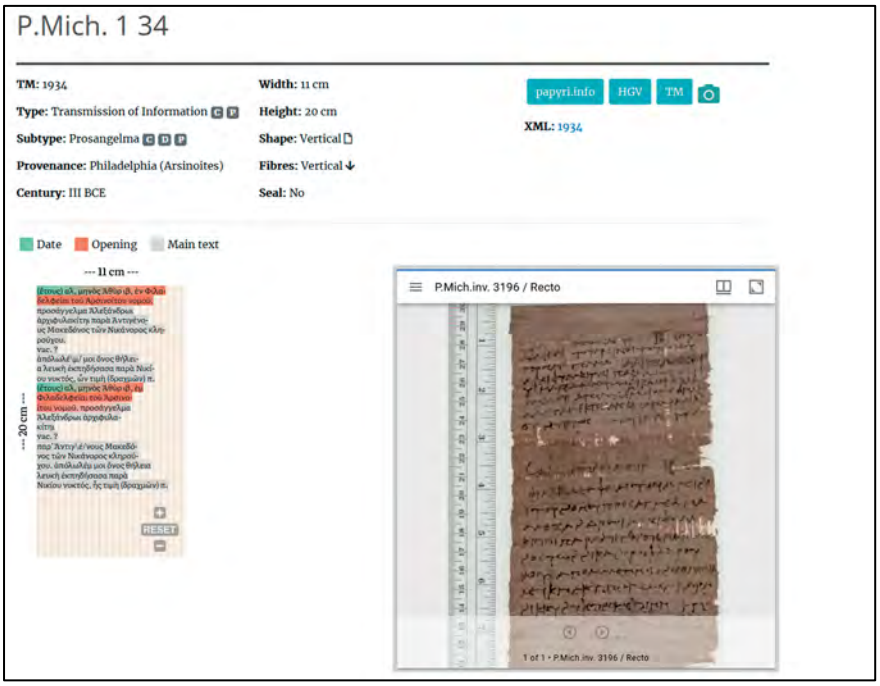


Fig. 1: P.Mich. I 34 (TM 1934) on *grammateus*. https://grammateus.unige.ch/doc/1934. Image: "P.Mich.inv. 3196; Recto". https://quod.lib.umich.edu/a/apis/x-1859/3196r_a.tif. University of Michigan Library Digital Collections. Accessed September 15, 2023.

22 Bonagura – Chang – Ferretti *et al.* 2023b, §§ 65–73.

3.2 Provenance and century

HGV is known for providing accurate information regarding both the provenance and datation of Greek papyri, openly available through Papyri.info and its API.[23] This is why we reuse their data for the date and place displayed on a *grammateus* page. A quick survey of the 63,435 HGV files accessed from Github on 2 June 2023 shows the encoding and the different ways the same information is given.

The date information is always encoded in `<origDate>` within the manuscript description (`<msDesc>`) of the TEI header. The date is given in plain text in German, with almost always the `@when`, `@notAfter` and `@notBefore` attributes for a date following the ISO standard. Those attributes may contain a full date (yyyy-mm-dd) or a less complete one (yyyy-mm, or yyyy only). The precision and certainty may be expressed with `@cert` or `@precision` attributes, or `<certainty>` and `<precision>` elements inside `<origDate>`. In the whole corpus, 2441 documents have more than one `<origDate>`, with the maximum being 5 different dates for P.Flor. II 185 (TM 11046).

```
1. <origDate when="-0113-08-26">26. Aug. 113 v.Chr.</origDate>
2. <origDate notBefore="-0275" notAfter="-0226" precision="low">Mitte III
   v.Chr.</origDate>
3. <origDate notBefore="0135-05" notAfter="0135-08">
   <precision degree="0.1"/>Sommer (?) 135
   </origDate>
4. <origDate notAfter="0201-03-20">
   <certainty locus="value" match="../month-from-date(@notAfter)"/>
   <certainty locus="value" match="../day-from-date(@notAfter)"/>
   <offset type="before" n="1">vor</offset> 20. März 201 (Monat und Tag un-
   sicher)</origDate>
```

Fig. 2: Four examples of HGV `<origDate>`. 1. BGU III 994 (TM 56); 2. PSI V 544 (TM 2166); 3. P.Oxy. LXXII 4878 (TM 114258); 4. SB XVI 12563 (TM 44381).

All documents have a provenance indication in German found in the `<origPlace>` element, which has no attribute.[24] The provenance may be completed further in `<placeName>` within the `<provenance>` element. Roughly 83 percent (52,846) of the documents have a `<provenance>` with at least one `<placeName>`. The ones without are mostly documents with an unknown place of origin. Again, there are variations in the encoding, and it is also worth noting the problematic related to places: the place may be

23 On the accuracy of HGV, see Cowey 1994; Reggiani 2017, 39–41; Cayless 2019, 38.

24 There is a single document that has the attributes `@evidence` and `@precision`. Only two documents did not have an `<origPlace>` element.

the place where the document was written, the place where it was sent to, the place it was found, or even its modern location.[25] The <placeName> information is more structured than the German text found in <origPlace>. Often the place is separated in different geographical levels: the settlement, nome and region. There are @ref attributes linking to the Pleiades gazetteer and to TM Places for disambiguation.[26] However not all place names are encoded with the same level of precision (compare Fig. 3 and 4).

```
1.  <origPlace>Tebetny bzw. Kerkeësis (Arsinoites)</origPlace>
2.  <provenance type="located">
    <p>
    <placeName type="ancient">Tebetny bzw. Kerkeësis (Arsinoites)
    </placeName>
    </p>
    </provenance>
```

Fig. 3: The HGV 1. <origPlace> and 2. <provenance> for SB XXVI 16825 (TM 11275).

```
    <provenance type="located">
    <p>
    <placeName type="ancient" ref="https://pleiades.stoa.org/places/737008
    https://www.trismegistos.org/place/1760">Philadelphia</placeName>
    <placeName type="ancient" subtype="nome" ref="https://www.trismegis-
    tos.org/place/332 https://pleiades.stoa.org/places/736893">Arsinoites
    </placeName>
    <placeName type="ancient" subtype="region">Ägypten</placeName>
    </p>
    </provenance>
```

Fig. 4: The HGV <provenance> for PSI V 544 (TM 2166).

In the *grammateus* example of TM 1934, the two metadata are retrieved from https://papyri.info/hgv/1934/source. The provenance shown on this page is simply the content of the <origPlace> element. The century is a simplification of the date given in the attributes of <origDate> elements. Our aim is to give a rough approximation and

25 See for example Sarri 2018, 59–60 for a discussion about the provenance of letters. In the HGV XML encoding, the different type of location may be given as a <provenance> @type with possible values being located, found, composed, sent, and acquired. The <placeName> @type may give the value either ancient or modern.

26 Pleiades: https://pleiades.stoa.org (accessed September 22, 2023), see Barker – Simon – Isaksen – de Soto Cañamares 2016; TM Places: https://www.trismegistos.org/geo/about.php (accessed September 22, 2023), for which see Verreth 2013.

estimate the administrative period – Ptolemaic, Roman or Byzantine – to which the text belongs. Some papyri have an incomplete date range, i.e. only one *terminus*, *post quem* (`@notAfter`) or *ante quem* (`@notBefore`): when that happens, we estimate a range of fifty years before or after the given terminus. When the date is not full in the format yyyy-mm-dd, we complete the missing information (always 01, which is admittedly arbitrary) to be able to calculate ranges. In case of multiple dates, we select the first in the list. Our aim is not to duplicate precious work that has already been done, and we refer users to HGV for the most precise information on date and place of origin.

While this dynamic access to Papyri.info allows us to display the latest state of knowledge regarding place and date, an obvious limitation is the impossibility of indexing the data for search purposes. To that end, we keep a copy of the HGV date and provenance locally which is regularly updated. It also serves as a backup in case Papyri.info is temporarily unavailable. This local data is used to create the filters of the search page.[27] For the provenance filter, for instance, we try to limit the options available to Nomes. However, not all documents have a `<placeName>` with a `@subtype` nome. The nome is still often given in parentheses in the `<origPlace>`. If we cannot safely extract a nome, we simply display the content of `<origPlace>`.

3.3 Texts

The case of textual content is slightly different in terms of how it is processed. We retrieve the transcription, for TM 1394 with the URL https://Papyri.info/ddbdp/p.mich;1;34/source. The XML is converted to HTML with the EpiDoc Stylesheets.[28] The HTML is then further transformed so that each line of text can be highlighted in colours corresponding to different text sections.

Because of the dynamic access, we cannot simply make a copy of the existing transcription and annotate it with our text sections at the precision level of our choice. We must rely on the existing EpiDoc tagging, and since words are not marked up, lines are the smallest textual units reliably accessible because they are consistently indicated with an `<lb/>` element. Thus, we include in our own metadata information about the text section pointing to line numbers.

```
<ab type="section">
<locus xml:id="section1" from="1" to="1" corresp="#m1" ana="../authority-
lists.xml#date"/>
<locus xml:id="section2" from="1" to="2" corresp="#m1" ana="../authority-
lists.xml#introduction"/>
```

27 https://grammateus.unige.ch/papyri.html (accessed September 13, 2023).

28 Eliott – Au – Bodard *et al.* 2008.

```
<locus xml:id="section3" from="3" to="10" corresp="#m1" ana="../authority-
lists.xml#main-text"/>
<locus xml:id="section4" from="11" to="11" corresp="#m1" ana="../authority-
lists.xml#date"/>
<locus xml:id="section5" from="11" to="13" corresp="#m1" ana="../authority-
lists.xml#introduction"/>
<locus xml:id="section6" from="13" to="21" corresp="#m1" ana="../authority-
lists.xml#main-text"/>
</ab>
```

Fig. 5: Text section encoding for TM 1934.

This encoding, with pointers to lines instead of words, explains why some lines have a gradation of colour as can be seen in Fig. 1: the first part of line 1 is a Date, followed by the beginning of the Opening in the second part of the line. Provided that there were never more than three sections on a line, we have found that this system is an acceptable compromise. In any case, the point of the visualisation is rather to show users which sections are present and in which order, not necessarily the exact point of start or end on a line.

As for date and place, dynamic access to the text prevents us from indexing and searching the content. But the same conceptualisation also applies: we are not interested in duplicating the remarkable search functions offered by Papyri.info and other projects. We do not keep a local copy of the text for search purposes, instead, we have included keywords in our metadata to help locate documents of a relevant type. The keywords are inspired from HGV subjects but are only available in English. They are usually composed of one word or a short expression.[29]

Some types of documents have been named differently in various languages, but even in English we can find variations: what we have called warrants have also been referred to as summonses and orders to arrest in the scientific literature. A keyword search for "warrant", "summons" or "order to arrest" would give exactly the same results. The keywords here serve to bridge the gap between a new classification and the expectations of papyrologists who are not familiar with our terminology.

Having reviewed *grammateus*' use of dynamic access to Papyri.info, the next section is devoted to the discussion of the issues we faced and how we addressed them. A central point to the dynamic access is the use of stable identifiers, which will be at the centre of the discussion.

29 Depending on when the XML was created, the keywords are broader in scope. For instance, business notes usually have several keywords about their content. Hence the single document PSI IV 407 (TM 2090) has the "painting" keyword, while most accounts simply have the "account" keyword.

4 Identifiers

While the use we make of place and date metadata is quite straightforward, dealing with the text is not so trivial. Since the purpose of the papyrus representation is to imitate the layout of the text as it is on the papyrus, it means that we need to know several parameters: the correct order of the text, but also what makes a single column, how many lines there are per columns, and how many characters per line. These are necessary to calculate how to fit the transcription inside the rectangle that is proportional to the papyrus. Moreover, we attempt to automatise the calculations as much as possible, to prevent the painstaking work of re-calculating by hand the font size for the hundreds of documents in our database. Although the process is satisfactory for the majority of our papyri, we still have a few special cases for which we need to intervene by hand with javascript code. Obtaining the above parameters is dependent on having the proper identifier of the correct transcription file and then, on making use of the EpiDoc encoding to determine what to show, calculate the font size and highlight the correct lines with their corresponding section colour.

4.1 Documents, Writing Surfaces, Texts, Editions

Papyrologists use an edition's publication title, volume and number to refer to papyri in academic writing, following the checklist of editions.[30] But papyri are subject to reedition, and these identifiers are therefore not stable by default. There are also different ways to think about papyri, depending on the focus of study: as texts (which may be separated in different documents), as documents (a physical object which may contain more than one text and may have survived as fragments), or a writing surface (one document may have several writing surfaces, but two separate physical objects are never considered one writing surface). Making the distinction between the three is not always straightforward, especially as documents could be recombined even in Antiquity: there are multi-layered documents with attachments, or the τόμος συγκολλήσιμος — rolls created by pasting together sheets of papyri.

Discussing the creation of Papyri.info, Hugh Cayless describes the difficulties to bring together identifiers originating from various projects, because these identifiers would not describe exactly the same thing.[31] While HGV identifiers are focussing on the text level, TM numbers are meant for a single writing surface, unless it is believed that the only relation between the texts is the writing surface itself.[32]

30 https://papyri.info/docs/checklist (accessed September 22, 2023).

31 Cayless 2019, 39. See also Reggiani 2017, 74–5 on the difference between the identifiers of HGV, TM and Papyri.info.

32 Depauw – Gheldof 2014.

As in *grammateus* we are interested in writing surfaces, the TM numbers are quite convenient to use, but there are still borderline cases where there is no TM number for the writing surface that we would like to show. For instance P.Oxy. IX 1212 (TM 28932) has what we have classified as a Warrant on the recto and a List on the verso, but it has a single TM number. This is likely because the two texts are of the same scribe, which counts as a relation other than the writing surface. In that case we have used the two HGV identifiers in *grammateus*, 28932a and 28932b.[33] PSI VII 807 (TM 17673) was classified as a Petition on the recto and an Account on the verso, however the texts are not distinguished with different HGV or other identifiers. Instead of inventing yet another identification scheme, we have set aside this document for the time being, as we have enough representative examples of both petitions and accounts.

While TM numbers cover in most cases the documents that we want to show, and are considered stable, they are not the identifier used to access either the HGV or DDbDP XML files.[34] As described above (p. 265), the identifiers for the transcriptions XML are the ddb-hybrid, created from the publication title. In the course of *grammateus*' four years of existence, we have seen examples of changed ddb-hybrid identifiers, such as the four papyri BGU II 472 Kol. I (TM 9196), P.Mich. II 121 V (11965), P.Mich. II 128 Kol. I a (11973) and P.Mich. V 238 r (12078) which have all had a change by the end of 2020.

There are two reasons when a ddb-hybrid identifier may change: if the text is reedited, the transcription is moved to another XML file, with the ddb-hybrid constructed from the new publication name. In other cases, the transcription can be split into two different XML files, as it happened to the four papyri listed above that were separated between recto and verso (TM 11965, 12078) or into columns (TM 9196, 11973).

If the identifier changed because of a reedition, it is still possible to maintain the dynamic access working, because the previous identifier normally still leads to an XML file containing a cross link to the new one.[35] PSI XV 1520 (TM 13778) and PSI XV 1528 (TM 14402) are examples of reedited papyri. For TM 13778, the split happened in 2013, and is documented in the Github history of the file,[36] whereas for TM 14402 both XML files have existed in parallel since the inception of Papyri.info's Github repository.[37] For these

33 This is the only case that we have had so far. Although this introduces a small inconsistency in the *grammateus* URLs, we have decided that it was worth it to show both examples in our database.

34 The only occasions when a TM number may disappear is if it is a double entry, or a fragment that was reunited. While we do not have an estimation of how many double entries may hide in TM, the chances of having a fragment reunited are low for *grammateus* documents, as we are working with complete documents (see p. 263 above).

35 See Cayless 2010 for a description of the redirection implementation.

36 The two ddb-hybrid identifiers are psi;15;1520 and psi.congr.xi;;7. See https://github.com/papyri/idp.data/commits/master/DDB_EpiDoc_XML/psi/psi.15/psi.15.1520.xml (accessed August 28, 2023).

37 The ddb-hybrid are psi;15;1528 and sb;12;11046. Because the reedition dates back to the 1970s, both files have a history going back to 2008: see https://github.com/papyri/idp.data/commits/master/DDB_EpiDoc_XML/psi/psi.15/psi.15.1528.xml and https://github.com/papyri/idp.data/commits/master/DDB_EpiDoc_XML/sb/sb.12/sb.12.11046.xml (Accessed August 28, 2023).

papyri, the XML with the previous edition's identifier has a `<ref>` element within the `<body>` of `@type` "reprint-in", and the `@n` attribute giving the updated identifier.[38] In such cases, it will be possible to implement a mechanism to check the presence of this element and extract the latest identifier from the attribute.

On the other hand, when the change occurs from a split transcription, then the previous XML files are not accessible anymore, and human intervention is necessary to maintain the database. This is a question to consider for the long-term sustainability of the project.

4.2 Recto, Verso, and other "textparts"

In their EpiDoc encodings, the texts are divided into `<div>`s of `@type` "textpart": these blocks of text may be the two sides, recto and verso, but also columns, fragments, marginal vs. central text, *scriptura interior* vs. *exterior*, and so on.

While discussing the three sets of identifiers, ddb, TM and HGV, I have already touched upon the question of sides, front and back – or recto and verso.[39] For all *grammateus* papyri, we need to display the text from a single side, and in most cases that side is the recto. The encoding is usually quite consistent, indicating as a `<div>` of `@type` "textpart" and `@n "r"` the recto, or `@n "v"` the verso. Therefore, it is possible to automatically hide the verso by hiding the textparts named as "v". However, there can always be exceptions: for instance, in P.Bagnall 46 (TM 219272), the columns are numbered with Roman instead of Arabic numerals, which led to the fifth column (numbered "v") not being displayed as it should. This is an example where the limits of automatization are reached: despite a very high level of accuracy there is no absolutely foolproof way to distinguish the recto or verso from other text components, and we had to intervene with javascript to correct the display. Another example is P.Oxy. XLII 3049 (TM 16447), where the verso is called Fragment B. We also intervene for P.Oxy. IX 1212 (TM 28932), the example noted above where we wanted to display both sides, and for the four papyri for which we display the verso instead of recto: P.Mich. II 121 V (TM 11965), P.Oxy. XII 1569 (31761), SB III 6262 (31055), and SB VI 9531 (25449). The latter two documents have the body of the letter written on the material verso, which was used as the front side by the scribe.

38 See https://papyri.info/ddbdp/psi.congr.xi;;7/source (TM 13778) and https://papyri.info/ddbdp/sb;12;11046/source (TM 14402) for instance (Accessed both August 28, 2023).

39 See Turner 1978 on the question of terminology and how recto and verso came to have a different meaning for papyri than codices. For codices, the recto was the side of the page that was written first, but at the start of the 20th century papyrologists started to apply the term recto to the inside of the roll and verso to the outside (cf. Turner 1978, 12). In *grammateus* we do not use the concepts of recto/verso but we adopted the front/back terminology to distinguish the order of writing. However, on Papyri.info the texts are encoded with the papyrological recto/verso distinction. Here the terms recto and verso will be used, as it is how sides are identified in the XML transcriptions.

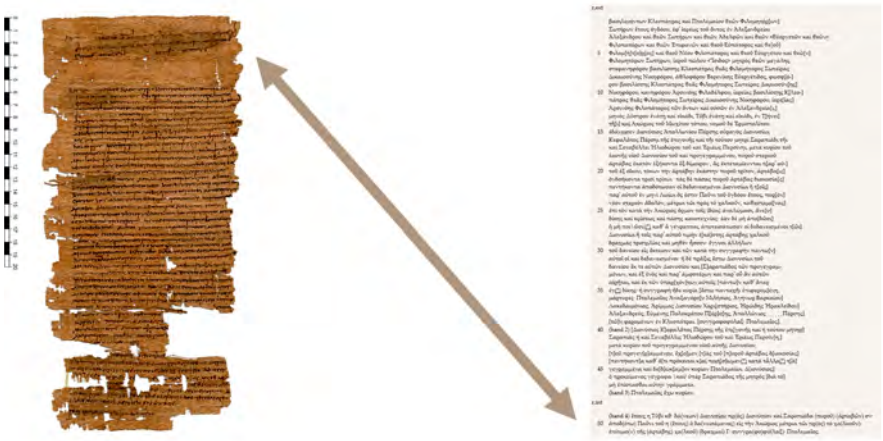


Fig. 6: P.Dion 16 (TM 3099) on the papyrus versus the Papyri.info transcription (https://papyri.info/ddbdp/p.dion;;16 accessed September 15, 2023). Image: Inv.Sorb. 2025 © Sorbonne Université – Institut de Papyrologie.

For *grammateus*, one important aspect is that we want to imitate the layout of a papyrus. However in some cases the transcription does not reflect how the text is actually laid out on the papyrus. As an example, compare the transcription and the image of P.Dion 16 (TM 3099), a six-witness document: its exterior text, the large paragraph, is the first in the transcription, and the interior text comes second, while it is the opposite on the papyrus. There is no obvious reason for an inverted transcription, except to match the edition and its line numbering: in the first edition of P.Rein. I 16, the interior script comes first but is not numbered.[40] In the second edition P.Dion. 16, the interior script is moved to the end of the exterior script and is numbered lines 49 to 51, maybe reflecting the order in which the text was written according to the editors.[41] That example is showing that what seems at first sight to be an inconsistency is in fact a valid encoding. While errors can be corrected on Papyri.info, we must remain very cautious before making any changes.

As for the recto and verso selection, it is often possible to recognize the interior script thanks to their textparts named with “int” or “ext”, and to force the display of the interior script first. However, it is again necessary to intervene, e.g. with PSI IV 379 (TM 2063) where interior and exterior are named fragments A and B according to the terminology adopted in the edition. Fragments A, B, etc. are not reliable to identify a textpart. As we have seen above, they may denote recto and verso, the interior and exterior script, or in other circumstances different columns.

40 Reinach – Ricci – Spiegelberg 1905, 89.

41 Boswinkel – Pestman 1982, 209–11.

4.3 Lines

What counts as a line? What does not? In the encoding documentation, it is stated that "[e]very line must have a line number".[42] What happens in practice, when a line is not numbered in the edition, is that it simply receives the same number as either the previous or following line. This is a fairly common phenomenon that we have observed in over 90 papyri from our database, but it complicates matters when it comes to identifying the exact lines of a text section to highlight in the papyrus display. Other inconsistencies concern papyri with multiple columns, where the line numbering may be continuous from start to finish or may start back at 1 for each column, e.g. BGU I 14 (TM 20191). When we started to encode text sections, we made the decision to count lines from start to finish regardless of how it was displayed in Papyri.info, but we counted only the same lines which were counted while ignoring the uncounted lines.[43]

When faced with an obvious error, we were able to correct the encoding through the Papyrological Editor: for example P.Wisc. II 80 (TM 13727) to separate two lines in two, or P.Köln. I 50 (TM 78) where the first line was entirely missing from the transcription. Other situations, while easily accommodated, have been discovered while checking our displays and noticing discrepancies in the text sections highlight: lines containing only seals, marks, or notes are usually not counted as far as we know.

Marks made on the papyrus probably by an official, such as the X on the first line of P.Oxy. VII 1028 (TM 20324), were identified in *grammateus* as a text section called "Official Mark".[44] Since the presence of an official mark was deemed relevant to the typology and highlighted as a text section, we had to be able to point to a line number. For TM 20324 or when marks are followed by text on the same line, the line gets numbered in the edition, see e.g. P.Oxy. XLIX 3520 and P.Oxy. XLIX 3516 (TM 15674 and 45256). But these marks are often not counted when they appear alone on a line, as in P.Oxy. XLIX 3514 or P.Oxy. XLIX 3515 (TM 45254 and 45255). When the official mark is at the beginning of the text on an uncounted line, we have counted it as line 0.[45] However, when the offi-cial mark is in the middle of the text, as for P.Oxy. XII 1452 (TM 21853) at the start of co-lumn 2, we need to count this line. It will create a discrepancy between our line counting and Papyri.info line counting, which we have tried to avoid.

For lines, as for all other aspects, respecting the edition is still important to the community, as can be seen for instance in the discussion regarding P.Count. 23 (TM 7767),

42 https://web.archive.org/web/20230418210852/https://papyri.info/docs/leiden_plus (accessed September 13, 2023).

43 It made sense at the time to match as closely as possible the line numbering from papyri.info, but if a newer edition is published where previously ignored lines are counted, as it happened to TM 3099, it will render our sections out of synchronisation with the latest publication.

44 Bonagura – Chang – Ferretti 2023b, § 72.

45 We also count line 1a of TM 114324 as line 0 for the column number text section, as lines with a letter are ignored in the line count but in this particular case, we need to highlight it.

showing the thoughtful discussion about line numbers and concern for matching the edition's numbering.[46] Line numbers are used in academic publications to cite the textual content of a papyrus, and readers should be able to find back the cited text, whether they consult the edition or Papyri.info. This means that line numbers may also incorporate any inconsistencies from printed editions that may follow different guidelines with regard to the layout, line numbering and other editorial features. However, we would like to stress that the inconsistencies, or rather heterogeneity of the Papyri.info XML transcriptions corpus is not a criticism. It is a phenomenon impossible to avoid in such a collaborative environment, and also given the longevity of the work and the successive transformations from SGML to TEI P5.[47] Uniformising the encoding while breaking the relationship with the printed edition would not be an improvement, and neither would be losing the collaborative aspect for the sake of consistency.

5 Conclusion

The dynamic access to data presented here was an important innovative aspect of *grammateus*: as far as we know, *grammateus* is the only project accessing data from Papyri.info in real time through its API while also adding a layer of information with the highlighted sections on top of the transcription. Therefore, we benefit from the up-to-date, peer-reviewed data on papyri. We are centralising the corrections efforts in a single online resource, instead of creating a separate database with transcriptions adapted to our need but also becoming obsolete as Papyri.info gets regularly updated. Centralising the efforts on a limited number of digital tools is not only a question of efficiency but also resources – financial and human as well. Both Papyri.info and Trismegistos must secure funding for their long-term sustainability: Trismegistos has introduced a subscription fee in 2020,[48] while Papyri.info launched in late 2019 a campaign to ensure the salary of a position for the day-to-day coordination of the interface.[49] As new tools are developed, it is clear that funding efforts of the community will concentrate on maintaining these already established and essential tools, particularly Papyri.info that serves as a central hub connecting existing resources and Trismegistos that provides so far the most stable identifiers. For that reason, being able to reuse Papyri.info material dynamically is all the more important to consider.

As Mark Depauw noted, Trismegistos Words is no longer synchronised with the latest texts available on Papyri.info but it would be desirable to be able to work directly

46 The discussion is available when clicking on "Editorial History" on Papyri.info: https://papyri.info/ddbdp/p.count;;23 (accessed September 13, 2023).

47 See Baumann – Bodard – Cayless *et al.* 2011 about the conversion to XML.

48 https://www.trismegistos.org/registration.php (accessed September 13, 2023).

49 https://www.supportpapyri.info (accessed September 13, 2023).

with the texts as they are.[50] To achieve this progress in the Linked Open Data environment for Greek papyri, he highlights the need for persistent identifiers at the word level, for other projects to point to words in the XML transcriptions. With the experiment of *grammateus* to do just that, to point to existing parts of the XML files, we have reached the same conclusion: it would be a great improvement to have persistent identifiers pointing not only to lines but also to blocks of text such as columns, or recto and verso sides. The presence of identifiers in parallel to existing line numbers, for instance, would allow to retain the existing line numbers of the edition for human readers, which may change with future re-editions, and it would also offer a more persistent mechanism for pointing to a specific line.

Whether these identifiers are added to the XML as `@xml:id` attributes, or whether it may be implemented through the nascent DTS API,[51] it is essential that we keep a format that the community can engage with relative ease as it is now the case with the XML and Leiden+ format available in the Papyrological Editor. Papyri.info is an incredibly valuable tool, in a great part because of the open collaborative aspect that lets all researchers participate directly and improve on the content. We hope that the example of *grammateus* can demonstrate that it is possible to work with a dynamic access to Papyri.info and how to improve its implementation through the use of identifiers.

Bibliography

Almas, B. – Cayless, H. – Clérice, T. – Jolivet, V. – Liuzzo, P. M. – Robie, J. – Romanello, M. – Scott, I. (2023), *Distributed Text Services (DTS): A Community-Built API to Publish and Consume Text Collections as Linked Data*. *Journal of the Text Encoding Initiative*, January, https://doi.org/10.4000/jtei.4352.

Babeu, A. (2011), *'Rome Wasn't Digitized in a Day': Building a Cyberinfrastructure for Digital Classics*. *Online*. Washington DC, https://doi.org/ISBN 978-1-932326-38-3.

Bagnall, R. S. (2010), *Integrated Digital Papyrology*, http://hdl.handle.net/2451/29592.

Barker, E. – Simon, R. – Isaksen, L. – Soto Cañamares, P. de. (2016), *The Pleiades Gazetteer and the Pelagios Project*, in *Placing Names: Enriching and Integrating Gazetteers*, ed. by M.L. Berman – R. Mostern – H. Southall, Bloomington, 97–109. [https://doi.org/10.2307/j.ctt2005zq7.12]

Baumann, R. (2013), *The Son of Suda On Line*, in *Digital Classicist 2013*, ed. by S. Dunn – S. Mahony, London, 91–106. [https://www.jstor.org/stable/44216325]

Baumann, R. – Bodard, G. – Cayless, H. – Sosin, J. – Viglianti, R. (2011), *Integrated Digital Papyrology*. Presented at the DH2011, Stanford, https://web.archive.org/web/20170710154928/http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-193.xml;brand=default.

Bonagura, G. – Chang, R.-L. – Ferretti, L. – Fogarty, S. – Nury, E. – Schubert, P. (2023a), *General Introduction*. Grammateus Project, https://grammateus.unige.ch/introduction/general.

Bonagura, G. – Chang, R.-L. – Ferretti, L. – Fogarty, S. – Nury, E. – Schubert, P. (2023b), *Key Concepts*. Grammateus Project, https://grammateus.unige.ch/introduction/concepts.

50 See M. Depauw's chapter in this volume.

51 Almas – Cayless – Clérice *et al.* 2023.

Boswinkel, E. – Pestmann, P. W., eds. (1982), *Les archives privées de Dionysios, fils de Kephalas*, Leiden.
Cayless, H. (2010), *Making a New Numbers Server for Papyri.Info*, Scriptio Continua (blog), 2 March 2010, https://philomousos.blogspot.com/2010/03/making-new-numbers-server-for.html.
Cayless, H. (2019), *Sustaining Linked Ancient World Data*, in *Digital Classical Philology*, ed. by M. Berti, Berlin – Boston, 35–50. [https://doi.org/10.1515/9783110599572-004]
Cowey, J. M. S. (1994), *Heidelberg Documentary Papyri Project*, in *Proceedings of the 20th International Congress of Papyrologists, Copenhagen, 23-29 August, 1992*, ed. by A. Bülow-Jacobsen, Copenhagen, 609–12.
Depauw, M. – Gheldof, T. (2014), *Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information*, in *Theory and Practice of Digital Libraries -- TPDL 2013 Selected Workshops*, ed. by Ł. Bolikowski – V. Casarosa – P. Goodale – N. Houssos – P. Manghi – J. Schirrwagen, Cham, 40–52. [https://doi.org/10.1007/978-3-319-08425-1_5]
Eliott, T. – Au, Z. – Bodard, G. – Cayless, H. – Lanz, C. – Lawrence, F. – Vanderbilt, S. – Viglianti, R. (2008), *EpiDoc Reference Stylesheets*, https://sourceforge.net/p/epidoc/wiki/Stylesheets/.
Ferretti, L. – Fogarty, S. – Nury, E. – Schubert, P. (2023a), *Description of Greek Documentary Papyri: Syngraphe*. Grammateus Project, https://doi.org/10.26037/yareta:zi3xod6ye5hivogrzg4dmxglqa.
Ferretti, L. – Fogarty, S. – Nury, E. – Schubert, P. (2023b), *Grammateus Project: Classification of Greek Documentary Papyri*. Grammateus Project, https://doi.org/10.26037/yareta:5mhltqzg2fdkrjat6miuud5bie.
Keersmaekers, A. – Depauw, M. – Broux, Y. (2016), *Trismegistos Words. A Database of Words in Greek Papyrological Texts*, https://www.trismegistos.org/words/about.php.
Manca Masciadri, M. – Montevecchi, O. (1984), *I contratti di baliatico*, Milan.
Montevecchi, O. (1988), *La papirologia*, 2nd ed., Milan.
Reggiani, N. (2017), *Digital Papyrology I*, Berlin – Boston.
Reinach, T. – Ricci, S. de – Spiegelberg, W. (1905), *Papyrus grecs et démotiques recueillis en Egypte*, Paris.
Riaño Rufilanchas, D. (2023), *Polyphemus, a Lexical Database of the Ancient Greek Papyri, and the Madrid Wordlist of Ancient Greek*. Presented at the DH2023, Graz.
Sarri, A. (2018), *Material Aspects of Letter Writing in the Graeco-Roman World, 500 BC-AD 300*, Berlin – Boston. [https://doi.org/10.1515/9783110426953]
Sosin, J. (2010), *Digital Papyrology*, The Stoa Consortium (blog), 26 October 2010, https://web.archive.org/web/20170408110552/http://www.stoa.org/archives/1263.
Turner, E. G. (1978), *The Terms Recto and Verso: the Anatomy of the Papyrus Roll*, Bruxelles.
Verreth, H. (2013), *A Survey of Toponyms in Egypt in the Graeco-Roman Period* (Trismegistos Online Publication 2), Leuven, https://www.trismegistos.org/top/.
Vierros, M. (2018), *Linguistic Annotation of the Digital Papyrological Corpus: Sematia*, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 105–18. [https://doi.org/10.1515/9783110547450-006]
Vierros, M. – Henriksson, E. (2017), *Preprocessing Greek Papyri for Linguistic Annotation*, https://hal.archives-ouvertes.fr/hal-01279493.

Marzia D'Angelo — Federica Nicolardi

Addressing Material Issues through Digital Solutions: *Maque-IT* and the Virtual Reconstruction of the Herculaneum Papyri

1 Overview

In spite of the destruction it caused, the eruption of Mount Vesuvius in 79 CE allowed a unique library to be preserved for seventeen centuries, by carbonizing and burying papyrus scrolls under a blanket of ashes, lapilli, and mud.[1] Nonetheless, the Herculaneum papyri, which were brought to light in the 18th century and are now stored in the *Officina dei Papiri Ercolanesi* (Biblioteca Nazionale di Napoli "Vittorio Emanuele III", Naples), are extremely fragile and complex. Due to their singular state and history of preservation, making the texts readable and accessible to scholars is only the last step in a long and complex process of material study of the artefacts that transmit them. In the past two decades, remarkable technological advancements have given rise to a "Digital Herculaneum Papyrology", which has been transforming the study of Herculaneum papyri and holds promise for further progress in the field of papyrology as a whole.[2]

One of the initial driving forces behind the digital approach undoubtedly originated from a crucial aspect for effective study of these ancient books, namely the need to enhance the legibility of the carbonized papyri, as the black ink is hard to distinguish from the black background. After several more or less successful experiments to effectively capture the papyri, starting from the late 19th century, mainly with the aim to accompany critical editions or to offer palaeographical *specimina*,[3] in the last two decades of the 20th century the photographic reproduction of the Herculaneum papyri began to

This paper is the result of close collaboration between the two authors, who take shared responsibility of the work described. As for the individual sections, the "Overview" and the section on "The needs for reconstruction" are by F. Nicolardi, while those on "Maque-IT" and "Future perspectives and *desiderata*" are by M. D'Angelo. We are very grateful to Nicola Reggiani for inviting us to the 2022 Digital Papyrology 3.0 Conference. This paper includes the description of new functionalities added to Maque-IT, presented at the RECREATE Hands on Reconstruction Training Workshop (Università degli Studi di Napoli Federico II, June 11–14, 2024).

1 For a general overview on the *Villa dei papiri* and Herculaneum papyrology see Longo Auricchio – Indelli – Leone – Del Mastro 2020.

2 See Fleischer 2021 and Reggiani 2021. Digital Herculaneum Papyrology is the subject of a specific section in the ENCODE open online course on the DariahTeach platform (D'Angelo, M. – Essler, H. – Nicolardi, F., *Unit 2.4. Digital Herculaneum Papyrology*).

3 Capasso 1983 and 1991, 142–8; Longo Auricchio – Indelli – Leone – Del Mastro 2020, 207–9.

https://doi.org/10.1515/9783111070162-017

respond specifically to the needs of study and to provide precise documentation.[4] The real turning point came when Steven Booras and David Seely of the *Center for the Preservation of Ancient Religious Texts* (Brigham Young University, Provo, Utah) applied multispectral photography technique to the Herculaneum papyri, imaging them in the infrared range and achieving excellent results.[5] The BYU imaging not only provided a valuable working tool with the precise aim of improving legibility and facilitating ink detection, but also inaugurated an era of extensive experimentation in the field of imaging for different purposes. From that moment onward, various techniques have been applied to the Herculaneum papyri, such as Reflectance Transformation Imaging (RTI),[6] shortwave-infrared (SWIR) within the hyperspectral range,[7] reflectography and X-ray fluorescence for ink analysis,[8] up to the ongoing activities aimed at producing 3D compilations of the opened scrolls combining multispectral and photogrammetric data[9] and the use of synchrotron radiation phase-contrast tomography and micro-CT aimed at virtual unwrapping of closed rolls.[10]

The advent of digital images providing millimetric detail also played a fundamental role in another respect, marking a substantial step forward in the development and implementation of methodologies for the virtual restoration of scrolls fragmented into different pieces.

2 The need for reconstruction

The Herculaneum scrolls' extremely precarious state of conservation, carbonized and fragmented into many thousands of pieces as they are, represents a major obstacle to reading them and to the systematic study of their texts. In most cases, reading a Herculaneum papyrus cannot be an immediate activity, but rather turns out to be a 'mediated'

4 The Norwegian team responsible for the unrolling of papyri from the mid-1980s (see Kleve – Angeli – Capasso *et al.* 1991) took two different kinds of photographs. The first consisted in the production of sequences of colour slides, to be viewed under a microscope. These reproductions, allowing letters and even minute traces of letters to be seen clearly, were the first to present themselves as a real working tool for the papyrologist. In addition to these, the Norwegian team also took a very large number of snapshots while working on opening papyri, to leave a record of their previous state and the original position of the various detached layers.

5 Booras – Seely 1999.

6 Piquette 2017.

7 Tournié – Fleischer – Bukreeva *et al.* 2019.

8 On ink analysis conducted on Herculaneum papyri see at least Seales 2009; Rabin – Schütz – Kohl *et al.* 2012; Rabin 2021; Brun – Cotte – Wright *et al.* 2016; Bonnerot – Del Mastro – Hammerstaedt *et al.* 2022.

9 Seales – Chapman – Nicolardi – Parker 2023.

10 Seales – Delattre 2013; Del Mastro – Delattre – Mocella 2015; Bukreeva – Mittone – Bravin *et al.* 2016; Stabile – Palermo – Bukreeva *et al.* 2021. On the successful use of virtual unwrapping on P.Herc.Paris. 4, see now Nicolardi – Parsons – Delattre *et al*. 2024 and https://scrollprize.org.

task. Far from being an end in itself, the technical task of virtually joining fragments and reconstructing a given scroll's format is a crucial preliminary step to recovering texts and contexts. Reconstructing a Herculaneum roll requires addressing a wide range of material issues, from those related to the eruption of 79 and the carbonization (alteration of the original cylindrical shape of the roll, the presence of fractures and other damages, reading difficulties due to charred surface, *etc.*) to those raised by the history of the opening and preservation of the scrolls.[11]

Significant material difficulties were caused by the two most widely used methods for opening and unrolling the papyri since the eighteenth century: (1) *scorzatura* ('peeling') and (2) *svolgimento* ('unrolling') by mechanical traction.

The first, particularly invasive technique consisted either in cutting the roll down the middle into two hemicylinders or in making two longitudinal cuts on each side of the papyrus, in order to separate the interior of the roll (*midollo*, i.e. marrow) from the outermost portions (*scorze*, i.e. barks). The stacks of *scorze* thus obtained were then put aside, obscuring the relationship between the *midollo* and the corresponding *scorze*, which were inventoried separately and later opened by scraping their layers away one at a time: once a layer had been transcribed, it was removed (destroyed) in order to reveal the layer below, until the last layer (*ultimo foglio*) was reached. The connection between *scorze* and *midolli* can only be determined through a scrupulous inspection of the whole collection, considering all the available data (script, bibliological data, content, and archive documentation). *Scorzatura* also resulted in the impossibility of obtaining consecutive columns from fragments of the same stack of *scorze* and the loss of information about the original position of the *scorza* in the roll. Thanks to geometrical calculations[12] and the use of virtual models (*maquette*s), papyrologists are now able to determine at least the relative position of the *scorze* and specify their distance in the virtual reconstruction of a scroll.

11 If one were to retrace some of the early milestones in the development of a methodology for reconstructing the Herculaneum scrolls, its embryonic phase could already be glimpsed in the productive period between the last quarter of the nineteenth century and the second half of the twentieth century, when numerous editions of Herculaneum papyri were published. Some of these editions, although impressive for their time, show a tendency to collect 'fragments' of dubious or completely unknown position, separating them from the better-preserved 'columns' with no particular concern to place them coherently within as complete a text as possible (e.g. Sudhaus 1892, 1895, 1896). In several cases, however, a precocious movement toward repositioning fragments can be discerned (e.g. Scotti 1839; Crönert 1900; Schober 1923). Awareness of the need for reconstruction has been gaining ground for several decades now, especially thanks to pioneering editions published by Obbink 1996, Janko 2000, 2010, 2020, Delattre 2007, Leone 2012. A revolutionary contribution was made by Essler 2008, who devised a highly effective geometric method making use of material and archival data. For an overview on recent developments in Herculaneum papyrology see D'Angelo – Essler – Nicolardi 2021.

12 Essler 2008. He also designed an easy-to-use spreadsheet collecting all the formulas needed for reconstruction (http://www.epikur-wuerzburg.de/downloads/MathRek.xls).

No less problematic is the reconstruction of papyri unrolled using the so-called *macchina di Piaggio*, the revolutionary machine invented by Father Antonio Piaggio in 1753. This machine worked by slowly detaching the outer layer from the underlying one, before it was reinforced with goldbeater's skin and then cut and fixed on a paper sheet. Nevertheless, the use of this ingenious technique often entailed irreparable damage to the papyrus surface and hindered reading of the text. Success in detaching the individual sheets was not always attained. On the contrary, it happened very often during this process that, due to the extreme cohesion of the scroll's circumferences (volutes), multiple layers of papyrus remained attached to one another. Borrowing a term from geology, scholars refer to this phenomenon as the scroll's 'stratigraphy'. A displaced fragment from a deeper circumference that has remained attached over the surface of an unrolled layer is called a *sovrapposto*; a displaced fragment torn off from an earlier circumference, remaining attached under the base layer, is called a *sottoposto* (Fig. 1).

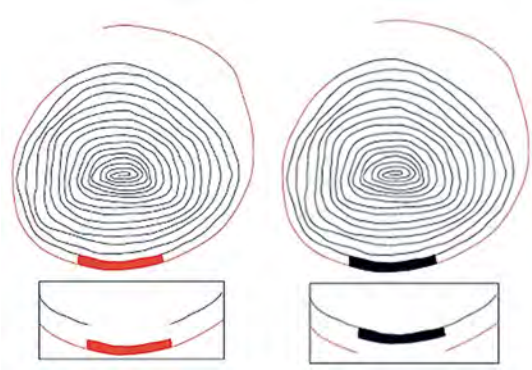


Fig. 1: Origin of a *sovrapposto* (on the left) and a *sottoposto* (on the right), from Nicolardi 2019.

Stratigraphy presents one of the most significant problems in the study of the Herculaneum papyri.[13] The presence of numerous displaced pieces can seriously compromise the reading and the understanding of a scroll's content, since the '*mise en colonne*' of the text appears totally out of order. Restoring the correct order of the layers in a virtual reconstruction is the only way to read the text. The necessary first step is to identify the displaced pieces by analyzing the papyrus under a microscope. Then, the *sovrapposti* have to be virtually placed in the later circumferences and the *sottoposti* in the earlier ones (Fig. 2).

13 The resulting displaced fragments were already occasionally identified and repositioned in the nineteenth century and at the beginning of the twentieth century, but we owe the first systematic approach to this problem to Nardelli 1973. Scholars have since started to reposition fragments more systematically, but papyri with more complex stratigraphy have generally been left aside.

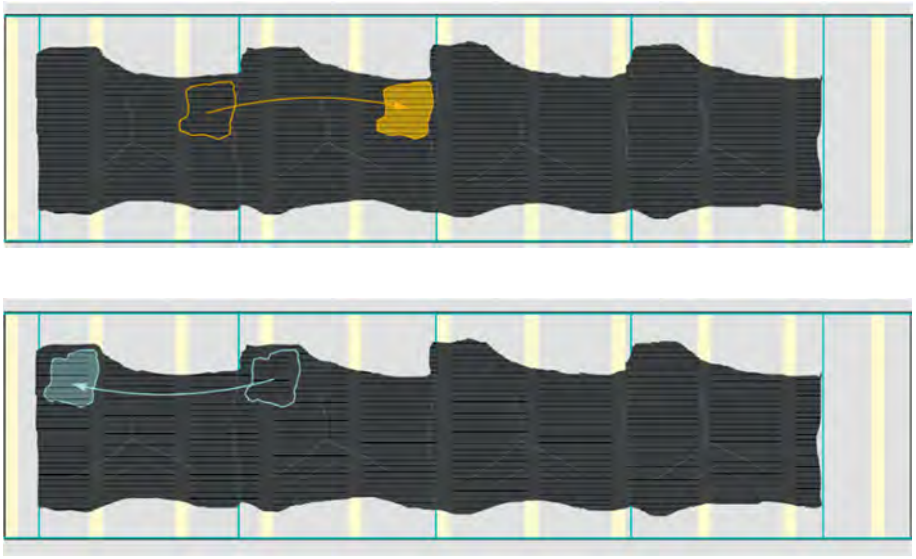


Fig. 2: Placement of a *sovrapposto* (top) and a *sottoposto* (bottom).

It is quite common for more than one layer to remain attached over or under the main surface. As a result, a first level *sovrapposto* must be moved one circumference forward, a second level *sottoposto* two circumferences back and so on. As is clear from Fig. 2, the occurrence of a *sovrapposto* necessarily results in a hole in the following circumference, just as a *sottoposto* results in a hole in the previous circumference. However, it is not always possible to observe this hole in the base layer, as such damages are often repeated in several consecutive circumferences and it is frequent that, after placing a *sovrapposto* or a *sottoposto*, one runs into another similar displaced fragment. This calls for a 'chain repositioning' (*spostamento a catena*).[14] Following some general rules, it is possible to 'predict' the stratigraphical state of a circumference based on the previous or the following ones, making the stratigraphical analysis of a piece more straightforward and certain. Correctly identifying and placing *sovrapposti* and *sottoposti* can revolutionize the study of a papyrus with complex stratigraphy, allowing papyrologists to reassemble lines and columns even from a papyrus that would otherwise only provide sequences of no more than a couple of letters.[15]

All these theoretical conclusions are perfectly suited for practical application leading to a smooth, self-monitored procedure.[16] At a more practical level, the need to repre-

14 Nicolardi 2019.

15 The first complete edition of a papyrus with extremely complex stratigraphy has recently been published by D'Angelo 2022.

16 Due to the special needs and peculiarity of the material, the reconstruction of the Herculaneum scrolls requires high specialization and attention to individual cases and particular features that may

sent the reconstruction in the form of a model was felt very early on by scholars concerned to restore the contents of Herculaneum *volumina*. This type of model, created by printing images, cutting them out, and pasting them on graph paper, was designed to reproduce the material features and layout of the scroll and to allow non-invasive restoration and repositioning of fragments. It is therefore, first and foremost, a tool for the editor of the text, but it can prove invaluable for readers as well. Absolutely pioneering was Delattre's choice to transfer his paper-based reconstruction model (the so-called *maquette*) of Philodemus' *De musica* to a digital medium (CD-ROM). The masterful edition of Epicurus' *De natura* book 2 by Giuliana Leone is also accompanied by a digital version of the paper *maquette*. A purely digital *maquette* was rendered for the first time in 2018[17] and the digital approach is increasingly establishing itself as the standard way of constructing these models.[18]

The structure of a *maquette* must be constructed by reproducing some basic data of the *volumen* relating to its *mise en page* and its material state:

– The model should reproduce the height of the columns (if known) and the average width of column + intercolumn, which represents the basic unit of the scroll as a book (comparable to a 'page');
– The width of the circumferences should be accurately reported in the model, taking into account the average regular decrease observed.

Since papyrologists have been looking for a tool to produce graphical renderings of their reconstructions, they have had to come to terms with the fact that no specialist software application is currently available to deal with the various aspects of reconstruction. It has therefore been necessary to adapt programs designed for completely different purposes, such as raster graphics editors or technical/architectural drawing tools. This renders the procedure for making the virtual *maquette* non-intuitive and time-consuming, and, despite the advantages of using a digital model, some understandably still prefer paper models.[19]

demand special treatment. This is the main reason why technical background information is often given in the introduction to the edition of a scroll, while the effort to collect data and systematically build up a general methodology from practical cases remains rare (e.g. Essler 2008; Janko 2016; D'Angelo – Nicolardi 2021; Nicolardi 2022). To date, the only essential 'guide' illustrating the steps involved in the edition of a Herculaneum roll, from reading and reconstructing on, is by Janko 2016.

17 Nicolardi 2018.

18 D'Angelo 2022.

19 See Janko 2020, 107: "I created a paper life-sized model of the entire roll, based on scaled prints of the digital images and on photocopies of the *disegni* where the originals do not survive; I find this easier to create and use than a digital model".

3 *Maque-IT*

The aforementioned reconstructing needs, with a particular focus on stratigraphical complexity, represented the basis for conceiving *Maque-IT*, the first image import and editing software tool specifically dedicated to the virtual reconstruction of papyri, with a particular focus on Herculaneum scrolls. The aim is to facilitate and speed up the complex work of repositioning the fragments and recomposing the original scroll format, providing the digital reconstructions with a higher degree of reliability. In particular, *Maque-IT* addresses two objectives: on the one hand, to automate the construction of the basic structure of the model of the roll; on the other, to enable papyrologists to check, and even to partially automate, their work with fragmentary texts.

Born in 2019 from an idea of the two authors, the project received over time different funding first to produce a prototype, developed in 2021 with the assistance of the computer scientist Florent Nöel (Curious Company), now to implement the fully functional and comprehensive version of the software, currently under development.[20] The current version of the program is a C++ desktop application using the Qt framework, which can be installed on both Windows and Mac operating systems.

At present, the software has two main functions: (1) it automatically creates the basic structure of the *maquette*; (2) it automatically moves displaced fragments (*sovrapposti* and *sottoposti*) of known provenance and prevents the user from moving a fragment to a position that is not valid according to the over-mentioned stratigraphical 'rules'.

Configuring the fundamental parameters (width of circumferences and decrease interval, width of column and intercolumn), *Maque-IT* is able to automatically build a basic model that can be extended and into which images of the papyrus fragments can be added. Although *Maque-IT* was initially designed to face the specific reconstruction challenges posed by the Herculaneum papyri, its functionality can be applied to any papyrus scroll by adjusting the basic parameters of the *volumen* under investigation. This makes it a versatile tool that can be profitably used by anyone working on the virtual reconstruction of any papyrus roll.

The main purpose behind the idea of *Maque-IT* was to provide scholars working on Herculaneum scrolls with complex stratigraphy with a tool that would facilitate this intricate operation and help them verify their work. Thus, from the very beginning, the software was developed to 'learn' the theoretical 'rules' of stratigraphy and apply them automatically. This tool cannot be a substitute for analysis of the papyrus by autopsy, but it makes the complex work of reordering pieces and layers in the virtual reconstruction easier and faster. This latter function has been tested on some papyri with complex

20 The development of the prototype was funded by the American Friends of Herculaneum and by the Centro Internazionale per lo Studio dei Papiri Ercolanesi 'Marcello Gigante'. The development of the final version is now planned within the project *RECREATE – REConstructing papyrus scrolls and REcovering Ancient TExts with the aid of a new digital tool* funded by the Fritz Thyssen Stiftung in 2023.

stratigraphy specifically selected as case studies. The reconstruction of one of these, *PHerc*. 89/1301/1383 (Philodemus, *[On gods]*),[21] will be used here as an example to illustrate the *Maque-IT* workflow.[22]

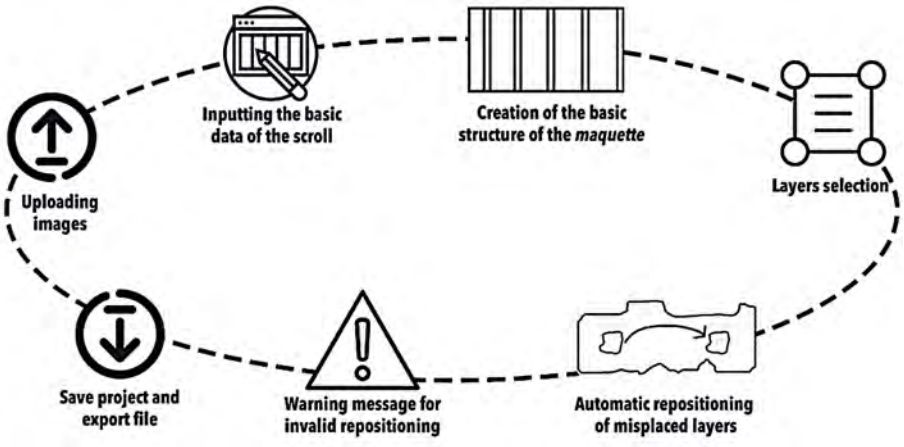


Fig. 3: *Maque-IT* workflow.

Uploading images. The first step is importing the images of the papyrus fragments.[23] In the current version, it is possible to import both images of individual fragments and images containing multiple fragments. A lasso tool allows the user to select and isolate individual fragments within the latter. When importing the first image, the user defines the scale of the papyrus. Further fragments can be scaled or adjusted to match the existing scale. Measurements of the imported fragments can be taken at any time using the ruler tool. To facilitate accurate measurements, all fragments are positioned on a millimeter grid.

Inputting the basic data of the scroll. Maque-IT simplifies the process of building a basic model of the scroll by automatically generating the succession of columns and decreasing circumferences, which has hitherto been done manually and is particularly time-consuming. By inputting the essential data obtained from the material analysis of the papyrus under investigation into the software, users can configure circumferences and columns. Measures in millimeters should be accurately taken on the original papy-

21 D'Angelo 2022.

22 For a general description of the software see also D'Angelo – Nicolardi 2021, 134–7. A demo video of the prototype is accessible on the website of the Centro Internazionale per lo Studio dei Papiri Ercolanesi (https://cispe.org/maque-it-un-software-per-la-ricostruzione-virtuale-dei-rotoli-ercolanesi-con-stratigrafia-complessa).

23 At the moment, high-resolution 2D images, but an experimentation with new 3D images is planned.

rus and provided in the software. To configure circumferences, users need to enter the width of a single circumference, the average width of the decreasing interval from the circumferences, and the desired number of circumferences. To configure the columns, users need to enter the average width of column + intercolumn (the basic unit of the scroll as a book), their height, and the desired number of columns.

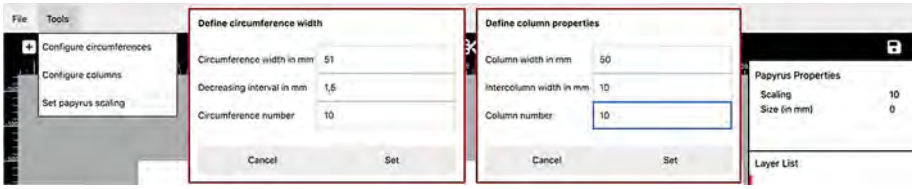


Fig. 4*: Maque-IT* Tools section.

Creation of the basic structure of the maquette. The circumferences are displayed as light blue outlined rectangles, each of them showing its width in mm at the top; the intercolumns are displayed as yellow rectangles. Once automatically created, circumferences and intercolumns can always be refined by manual adjustments. When the width of a circumference is manually modified, the measurement at the top will also be automatically updated. The basic scroll model can be extended at any time by adding more circumferences and columns to the beginning and end.

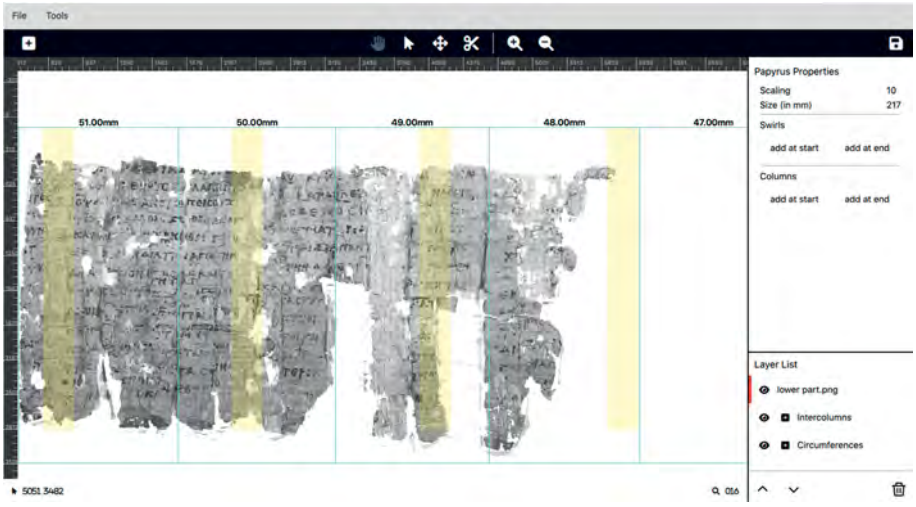
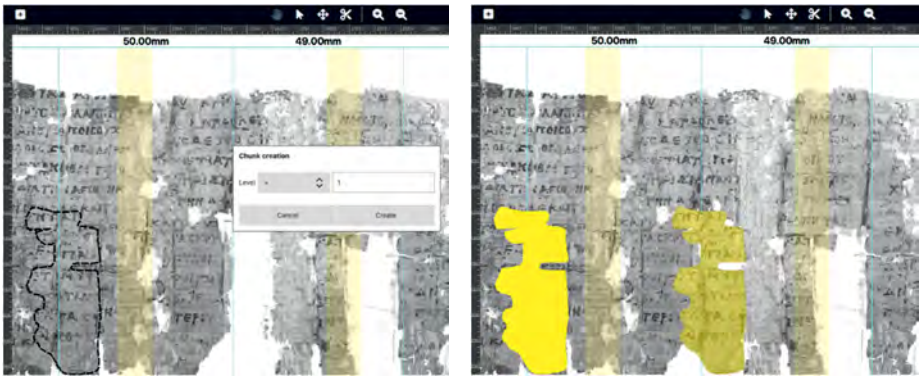


Fig. 5*: Maque-IT* basic model.

Layers selection. The software provides a selection tool that enables recording on the digital image of a Herculaneum papyrus the stratigraphical analysis carried out under a microscope. Users can select regions corresponding to displaced layers and add a tag (and a related colored outline) indicating whether they are *sovrapposti* or *sottoposti* and of what level (1+, 2+, 3+, 4+ …; 1-, 2-, 3-, 4- …).

Automatic repositioning of misplaced layers. After completing the selection, the software automatically moves the misplaced layers one or more circumferences back or forward, depending on the added tag (e.g., '1+' moves circumference forward, '2-' moves two circumferences back, and so on …), according to the established rules. In the images below, the area corresponding to a first-level *sovrapposto*, after being selected (Fig. 6), has been automatically moved one circumference forward (Fig. 7). In the model, the misplaced portions appear fully colored (with different colors assigned to different levels of *sovrapposti* and *sottoposti*), while the repositioned ones have the same color but with transparency.



Figs. 6–7: On the left, selection of a layer and addition of the related tag (here, first-level *sovrapposto* = 1+); on the right, automatic repositioning of the layer one circumference forward.

Warning message for invalid repositioning. To ensure repositioning accuracy, *Maque-IT* prevents the user, via an error message, from moving a layer to a position that is not valid according to the stratigraphical rules. For instance, one of the fundamental rules of the 'chain repositioning' of *sovrapposti* and *sottoposti* is that a *sovrapposto-sottoposto* sequence cannot occur.[24] If the user identifies a layer that breaks the stratigraphical rules, a warning message appears, informing that the move is invalid and suggesting a level change (Fig. 8).

Save project and export file. *Maque-IT* allows to save the current project as an editable file and export the reconstruction in .tiff format at any time.

24 See Nicolardi 2019.

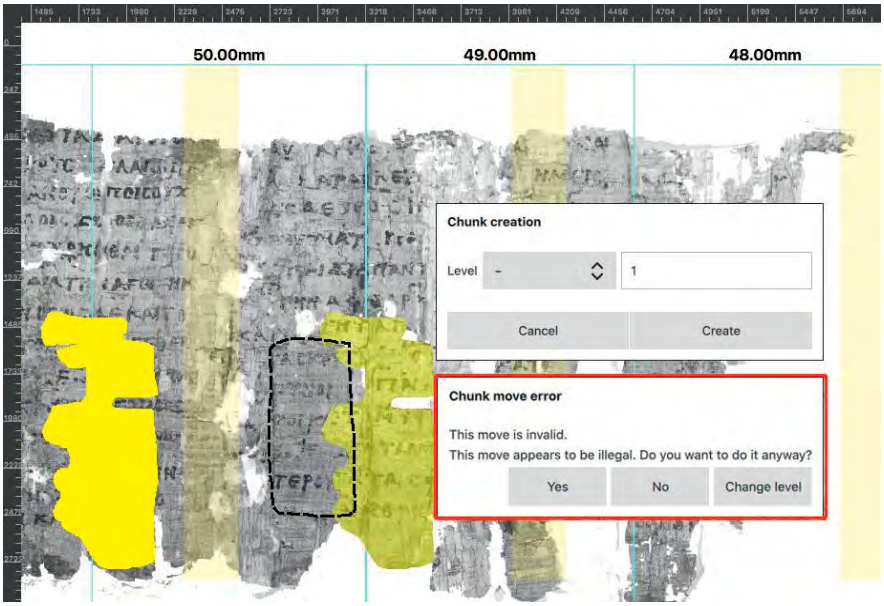


Fig. 8: Warning message for invalid repositioning.

Correctly identifying and placing *sovrapposti* and *sottoposti* can revolutionize the study of papyri with complex stratigraphy, enabling papyrologists to reassemble lines and columns even from a papyrus that would otherwise only provide sequences of no more than just a couple of letters[25]. In such cases, repositioning the misplaced layers allows not only the restoration of the original succession of the identifiable text columns but also the recovery of new text columns, reconstructed for the first time through the association of letter sequences previously scattered across different levels. In such cases, a final confirmation of the correct replacement can be provided by textual data (Fig. 9).

4 Future perspectives and *desiderata*

The current version of *Maque-IT* is still a prototype that requires improvement and implementation, first of all to ensure stability when dealing with a high number of images, as when rendering the reconstruction of an entire scroll, as well as fluency and smooth user experience. Paying attention to these aspects will entail adding new functionalities that can make the tool helpful for any papyrologist working on the virtual restoration of a scroll with known parameters.

25 D'Angelo 2022, part. 86–90, 108–11.

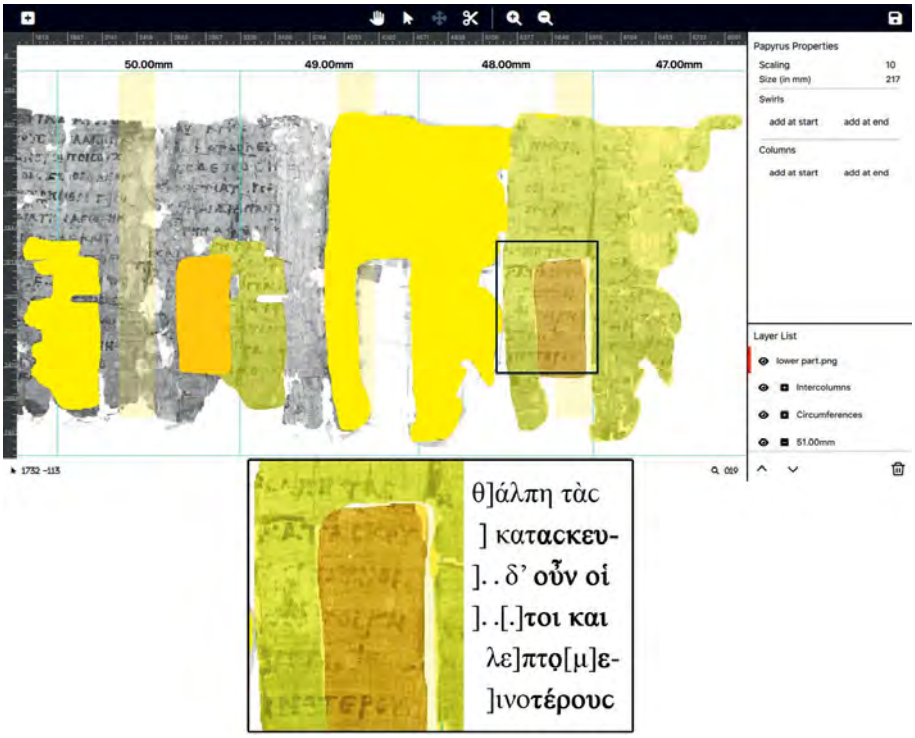


Fig. 9: Reconstruction carried out using *Maque-IT*. At the bottom, detail of the text recovered after repositioning two *sovrapposti* (Philodemus, *[On Gods]*, col. 151, 32-37 D'Angelo 2022).

While remaining open to implementing features that cater to different reconstructing needs, the primary focus of *Maque-IT* is to become a comprehensive tool for the reconstruction of the Herculaneum papyri, promoting remote and non-invasive study of the collection. The final version aims to include everything needed for the recomposition of a carbonized scroll, even new functions allowing for easy and intuitive calculations. In particular, it will be possible to estimate the total length of the original scroll by using the available material, bibliological, and archival data. A dedicated new section is planned for the reconstruction of scrolls subjected to *scorzatura*, aimed at automatically placing the *scorze* in the reconstruction model according to the circumference calculated from input data. This section will support users in applying reconstructing geometric methods currently in use, benefiting greatly from the integration of MathRek, an easy-to-use spreadsheet designed by Holger Essler that collects all the formulas needed for reconstruction.[26] As far as the stratigraphical issues, the software will become smart-

26 See above, nn. 10 and 11.

er and learn how to predict the presence of displaced layers, suggesting their possible placement according to the rules of 'chain repositioning'.

At a later stage of its development, *Maque-IT* will explore some completely novel and hitherto unexplored possibilities. The first step will be to evaluate the potential benefits of using the new 3D images of the Herculaneum papyri.[27] Another *desideratum* is to enable the inclusion of the facsimiles (*disegni*) of the fragments alongside the images of the papyrus. This could prove particularly helpful in cases where they display text no longer preserved on the original papyrus. Finally, a new text-image alignment function will be tested, allowing users to add parallel transcriptions of the texts while they work on the reconstruction. This feature will provide users with the opportunity to work with both the visual representation and the transcribed content simultaneously.

5 Bibliography

Bonnerot, O. – Del Mastro, G. – Hammerstaedt, J. – Mocella, V. – Rabin, I. (2022), *XRF Ink Analysis of Selected Fragments from the Herculaneum Collection of the Biblioteca Nazionale di Napoli*, in *Proceedings of the 29th International Congress of Papyrology*, ed. by M. Capasso – P. Davoli – N. Pellé, Lecce, 200–13.

Booras, S.W. – Seely, D.R. (1999), *Multispectral Imaging of the Herculaneum Papyri*, CErc 29, 95–100.

Brun, E. – Cotte, M. – Wright, J. – Ruat, M. – Tack, P. – Vincze, L. – Ferrero, C. – Delattre, D. – Mocella, V. (2016), *Revealing Metallic Ink in Herculaneum Papyri*, in *Proceedings of the National Academy of Sciences of the United States of America*, 113,14, 3751–4, doi:10.1073/pnas.1519958113.

Bukreeva, I. – Mittone, A. – Bravin, A. – Festa, G. – Alessandrelli, M. – Coan, P. – Formoso, V. – Agostino, R. G. – Giocondo, M. – Ciuchi, F. – Fratini, M. – Massimi, L. – Lamarra, A. – Andreani, C. – Bartolino, R. – Gigli, G. – Ranocchia, G. – Cedola, A. (2016), *Virtual Unrolling and Deciphering of Herculaneum Papyri by X-Ray Phase-Contrast Tomography*, Scientific Reports 6, 27227, 2016, doi: 10.1038/srep27227.

Capasso, M. (1983), *Storia fotografica dell'Officina dei Papiri Ercolanesi*, Napoli.

Capasso, M. (1991), *Manuale di Papirologia ercolanese*, Galatina.

Crönert, W. (1900), *Der Epikureer Philonides*, Sitzungsbericht der Königlich Preußischen Akademie der Wissenschaften zu Berlin 2, 942-59.

D'Angelo, M. (2022), *Filodemo. Opera incerta sugli dèi*, Napoli.

D'Angelo, M. – Essler, H. – Nicolardi, F. (2021), curr., *Tracing the Same Path. Tradizione e innovazione nella papirologia ercolanese/Tradition und Fortschritt der herkulanischen Papyrologie zwischen Deutschland und Italien*, Napoli.

D'Angelo, M. – Nicolardi, F. (2021), *Dalla ricostruzione all'edizione dei papiri ercolanesi: problemi e proposte di presentazione e rappresentazione*, in D'Angelo – Essler – Nicolardi 2021, 121–38.

Delattre, D. (2007), *Philodème de Gadara, Sur la musique, livre IV*, Paris.

Del Mastro, G. – Delattre, D. – Mocella, V. (2015), *Una nuova tecnologia per la lettura non invasiva dei papiri ercolanesi*, CErc 45, 227–30.

Essler, H. (2008), *Rekonstruktion von Papyrusrollen auf mathematischer Grundlage*, CErc 38, 273–307.

Fleischer, K. (2021), *Die Papyri Herkulaneums im Digitalen Zeitalter: Neue Texte durch neue Techniken – eine Kurzeinführung*, Berlin – Boston.

Janko, R. (2000), *Philodemus, On Poems, Book One*, Oxford – New York.

27 See above and n. 8.

Janko, R. (2010), *Philodemus, On Poems, Books 3-4*, Oxford – New York.

Janko, R. (2016), *How to Read and Reconstruct a Herculaneum Papyrus*, in *Ars Edendi Lecture Series* IV, ed. by B. Crostini – G. Iversen – B. M. Jensen, Stockholm, 117–61. DOI: http://dx.doi.org/10.16993/baj.f.

Janko, R. (2020), *Philodemus, On Poems, Book 2: With the fragments of Heracleodorus and Pausimachus*, Oxford – New York.

Kleve, K. – Angeli, A. – Capasso, M. – Fosse, B. – Jensens, R. – Störmer, F. C. (1991), *Three Technical Guides to the Papyri of Herculaneum. How to Unroll. How to Remove Sovrapposti. How to Take Pictures*, CErc 21, 111–24.

Leone, G. (2012), *Epicuro, Sulla natura, libro II*, Napoli.

Longo Auricchio, F. – Indelli, G. – Leone, G. – Del Mastro, G. (2020), *La villa dei papiri. Una residenza antica e la sua biblioteca*, Roma.

Nardelli, M.L. (1973), *Ripristino topografico di sovrapposti e sottoposti in alcuni Papiri Ercolanesi*, CErc 3, 104–15.

Nicolardi, F. (2019), *Aspetti e problemi della stratigrafia nei papiri ercolanesi: lo spostamento a catena di sovrapposti e sottoposti*, CErc 49, 191–216.

Nicolardi, F. (2022), *Per la ricomposizione di rotoli ercolanesi scorzati*, in *Proceedings of the 29th International Congress of Papyrology*, ed. by M. Capasso – P. Davoli – N. Pellé, Lecce, II, 751–8.

Nicolardi, F. – Parsons, S. – Delattre, D. – Del Mastro, G. – Fowler, R. – Janko, R. – Reinhardt, T. – Parker, C. S. – Chapman, C. – Seales, W. B. (2024), *Revealing Text from a Still-Rolled Herculaneum Papyrus Scroll (PHerc.Paris. 4)*, ZPE 229, 1–13 (available at https://www.habelt.de/openaccess).

Obbink, D. (1996), *Philodemus, On Piety, Part 1, critical text with commentary*, Oxford.

Piquette, K. (2017), *Illuminating the Herculaneum Papyri: Testing New Imaging Techniques on Unrolled Carbonised Manuscript Fragments*, Digital Classics Online 3.2, 80–102.

Rabin, I. (2021), *Inchiostri nell'Antichità*, in D'Angelo – Essler – Nicolardi 2021, 175–80.

Rabin, I. – Schütz, R. – Kohl, A. – Wolff, T. – Tagle, R. – Pentzien, S. – Hahn, O. – Emmel, S. (2012), *Identification and Classification of Historical Writing Inks in Spectroscopy*, COMSt Newsletter 3, 26–30.

Reggiani, N. (2021), *I rotoli di Ercolano, la papirologia virtuale e l'edizione critica digitale dei papiri: alcune riflessioni*, in D'Angelo – Essler – Nicolardi 2021, 163–7.

Scotti, A. A. (1839), *Herculanensium Voluminum Quae supersunt tomus VI, PHerc. 152/157*, Neapoli.

Schober, A. (1923), *Philodemi De pietate Pars prior*, Regiomont, diss., CErc 18 (1988), 67–125.

Seales, W. B. (2009), *Lire sans détruire les papyrus carbonisés d'Herculanum*, Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres 153/2, 907–23.

Seales, W. B. – Chapman, C. – Nicolardi, F. – Parker, C. S. (2023), *The Digital Restoration of the Herculaneum Papyri*, CErc 53, 201–11.

Seales, W.B. – Delattre, D. (2013), *Virtual Unrolling of Carbonized Herculaneum Scrolls: Research Status (2007-2012)*, CErc 43, 191–208.

Stabile, S. – Palermo, F. – Bukreeva, I. – Mele, D. – Formoso, V. – Bartolino, R. – Cedola, A. (2021), *Computational platform for the virtual unfolding of Herculaneum Papyri*, Scientific Reports 11, 1695, https://doi.org/10.1038/s41598-020-80458-z.

Sudhaus, S. (1892), *Philodemi volumina rhetorica*, Lipsiae.

Sudhaus, S. (1895), *Philodemi volumina rhetorica. Supplementum*, Lipsiae.

Sudhaus, S. (1896), *Philodemi volumina rhetorica II*, Lipsiae.

Tournié, A. – Fleischer, K. – Bukreeva, I. – Palermo, F. – Perino, M. – Cedola, A. – Andraud, C. – Ranocchia, G. (2019), *Ancient Greek Text Concealed on the Back of Unrolled Papyrus Revealed through Shortwave-infrared Hyperspectral Imaging*, Science Advances, 5,10, eaav8936.

Vincenzo Damiani

Automated Layout Segmentation and Text Recognition for Literary Papyri and Incunabula

A Case Report from the *Anagnosis* Project

1 Introduction

This contribution bears some similarities to that of Federica Nicolardi and Marzia D'Angelo in this same volume in that the *Anagnosis* project, which will be described in detail in the following pages, was initially designed to address a gap in the field of Herculaneum papyrology.

Anagnosis was integrated into the interdisciplinary *Kallimachos* project, funded by the German Ministry of Education and Research and headquartered at the University of Würzburg. *Kallimachos* brought together diverse research disciplines, including papyrology, German literature, local history, and philosophy. While some sub-projects were permanently integrated into a *Centre for Digital Philology* (*Zentrum für Philologie und Digitalität*)[1] after the main project *Kallimachos* ended, *Anagnosis* was unfortunately not able to benefit from this integration. However, the questions initially posed by *Anagnosis* can still be relevant and may be addressed again in the future, potentially within a different format and framework.

I had the opportunity to work in close collaboration with Holger Essler (Würzburg / Venice) on *Anagnosis*, with him playing a key role in developing the original scope and, in conjunction with Michael Erler (Würzburg), supervising the project as a whole. Holger Essler and I were responsible for overseeing the implementation of the software that I will outline later. The *Anagnosis* project was fortunate to receive funding in two phases (2014–2017; 2017–2019), allowing us to focus on different tasks during each phase. During the first phase, we concentrated on conducting a digital analysis of Greek literary papyri, while during the second phase, we dedicated our efforts to analysing the first Greek texts that were printed at Aldo Manuzio's (ca. 1452–1515) workshop.

1 https://www.uni-wuerzburg.de/zpd/startseite. All hyperlinks last accessed on 25.6.2024.

https://doi.org/10.1515/9783111070162-018

2 *Anagnosis* I

A summary of the objectives, accomplishments, and materials utilised during the first phase of the *Anagnosis* project has already been provided by Holger Essler in a contribution (co-authored with Rodney Ast [Heidelberg]) that was published in *Digital Papyrology II*, edited by Nicola Reggiani, in 2018.[2] In this current paper, instead of reiterating this information, I will primarily delve into the second phase of the project which has not yet been discussed in previous publications.

In the first phase (2014–2017), our focus was on digitally editing Greek literary papyri, particularly those from Herculaneum. One of our main goals was to improve and enhance the encoding of Herculaneum texts for the *Digital Corpus of Literary Papyri* (DCLP).[3] This encoding process is now largely complete. The full texts can be accessed on papyri.info or directly on the website of the *Würzburger Zentrum für Epikureismusforschung* as part of the *Thesaurus Herculanensium Voluminum*.[4] From 2018 until the present, the encoding of papyri, not only from Herculaneum, has continued (partly in parallel with *Anagnosis*) through the EKDOSIS (completed in 2020),[5] *Encode*,[6] and *E(dendo)discimus*[7] initiatives. These are all part of the *Editio Maior* initiative.

While encoding texts for the *Digital Corpus of Literary Papyri*, we began to develop a web-based tool for automated layout analysis. Many of the texts we had encoded were used as part of the dataset for this tool. The aim was to align digital transcriptions with images of literary papyri from various collections worldwide. This alignment of transcribed text and images had two main goals: 1) on the one hand, to make it easier for users without specialised palaeographical knowledge to read the original documents; 2) on the other hand, to extract individual letters from each fragment. Letter extraction had a twofold purpose. 2.a.) Firstly, automatically creating alphabets for each fragment would have contributed to the palaeographical analysis, for example, by grouping together similar handwriting styles; 2.b.) secondly, using these alphabets to – virtually and visually – reconstruct lacunae would have been helpful in determining the feasibility of different text integrations in the first place. It is important to note that, at least in this initial stage, the *Anagnosis* project did *not* aim to implement OCR for handwritten text but rather to align images and existing transcriptions.

Let me provide a brief sketch of how the *Anagnosis* editor functioned. The software is no longer available on the *Kallimachos* server but can be retrieved as a local version upon request. After logging in, the editor allows users to open a new document by enter-

2 Ast – Essler 2018.
3 https://papyri.info/browse/dclp.
4 https://epikur-wuerzburg.de/aktivitaeten/editio/thv.
5 https://epikur-wuerzburg.de/aktivitaeten/editio/ekdosis.
6 https://epikur-wuerzburg.de/aktivitaeten/editio/encode.
7 https://epikur-wuerzburg.de/aktivitaeten/editio/e-discimus.

ing the TM number of the papyrus (Fig. 1). From there, users can navigate through the transcription and select a fragment or column of text (Fig. 2). When a text unit is selected, the corresponding image automatically appears on the left side of the screen, linked to the transcription through the `corresp`-attribute in the XML version of the encoded text.[8] If no image is available, users can manually add one and provide corresponding metadata. To perform layout analysis and letter matching, users have to manually mark the writing area on the papyrus (norm box), which serves as a reference for the coordinates of line boxes and character boxes within it. (Fig. 3) As a result of the alignment, hovering the mouse cursor over the image or transcription highlights the corresponding character. The first version of the tool only worked with handwritten reproductions of Herculaneum Papyri (the so-called *disegni* or the engravings taken from them) (Fig. 4).

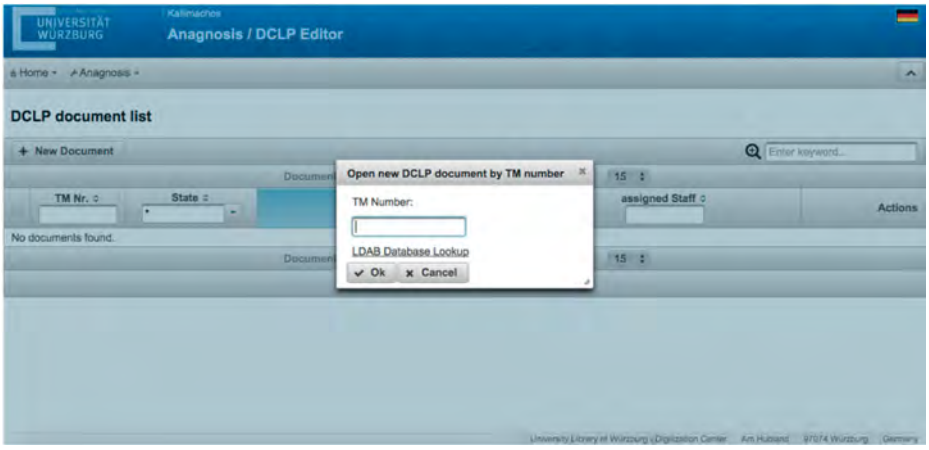


Fig. 1: *Anagnosis* editor, opening new DCLP document by TM number.

8 Ast – Essler 2018, 70.

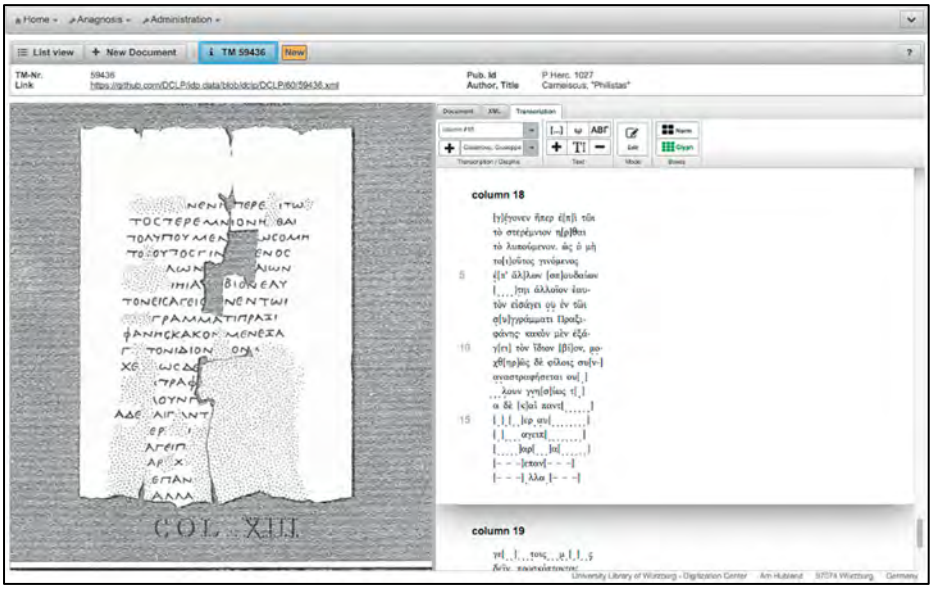


Fig. 2: *Anagnosis* editor, selecting and displaying a column of text.

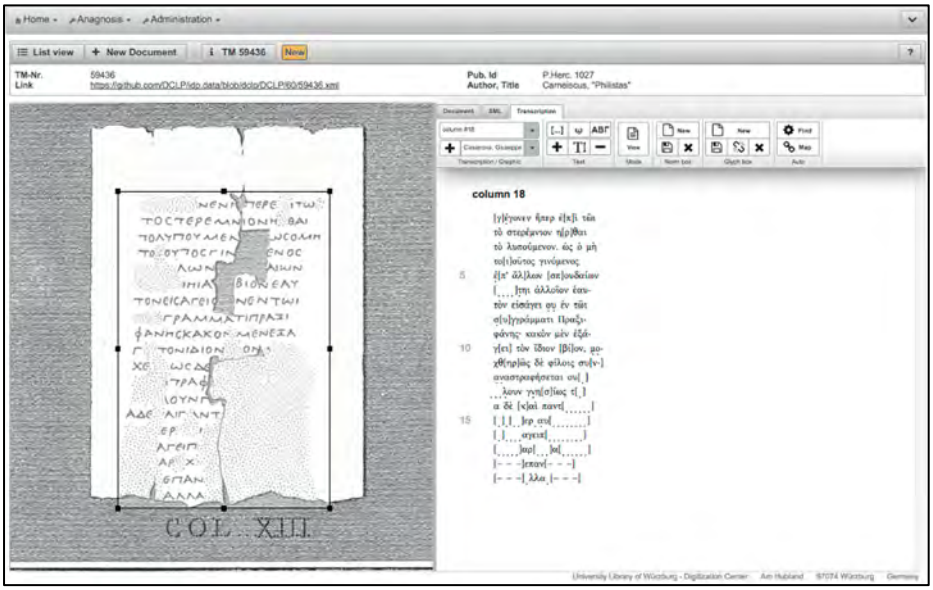


Fig. 3: *Anagnosis* editor, marking the writing area on the papyrus.

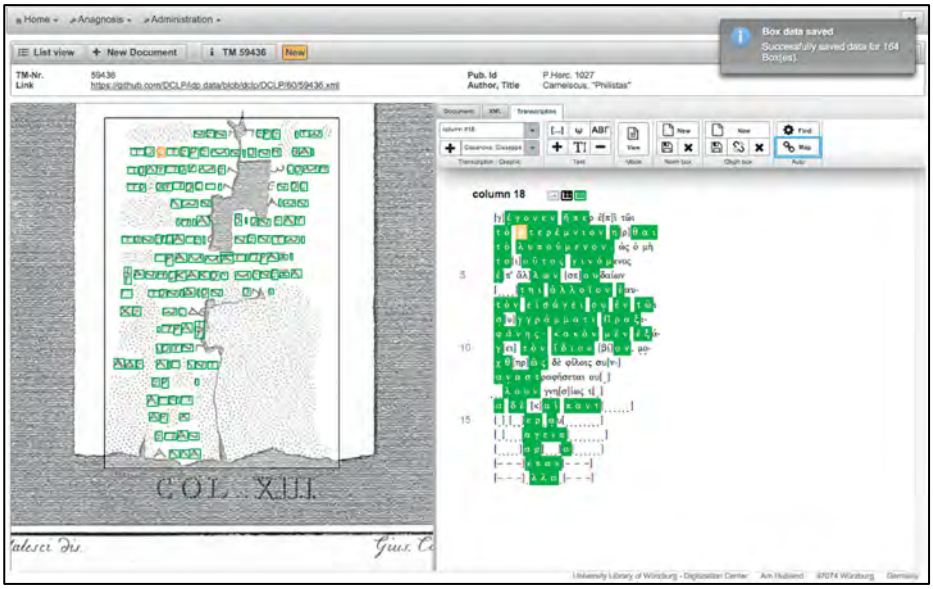


Fig. 4: *Anagnosis* editor, text-image alignment.

A second version of the *Anagnosis* editor, developed in partnership with the *German Research Centre for Artificial Intelligence* in Kaiserslautern, used Computer Vision techniques called *Dense Scale Invariant Feature Transform* (Dense SIFT) and *Four-Patch Local Binary Pattern* (FP-LBP) to detect text lines and individual characters in real photographs (Fig. 5). Dense SIFT is a method that extracts local features from an image by densely sampling key points in a regular grid across the entire image. These key points are then described using SIFT (Scale Invariant Feature Transform) descriptors, which capture the local appearance and orientation of the key points in the image.[9] FPLBP, on the other hand, is a texture descriptor that encodes the spatial relationship between the pixels in an image by comparing the intensities of four neighbouring pixels.[10] While this tool component demonstrated great potential, it remained in prototype status. Other projects, such as the one presented at the International Papyrology Congress in Lecce in 2019 by Benjamin Kiessling, Daniel Stökl, Rodney Ast, and Holger Essler,[11] have also aimed to align existing images and transcripts for both documentary and literary papyri.

9 Lowe 1999; 2004.
10 Ojala – Pietikäinen – Harwood 1994 and 1996.
11 Kiessling – Stökl Ben Ezra – Ast – Essler 2019.

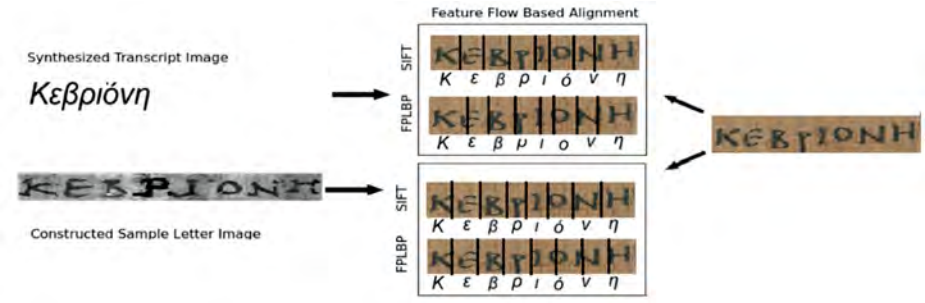


Fig. 5: Text line and character recognition through Dense SIFT and FP-LBP.

3 *Anagnosis* II

During the second phase of *Anagnosis* (2017–2019), the focus was still on automatic layout analysis but with a different set of documents. Rather than working with literary papyri, we focused on early printed versions of Greek texts from the Venice printing house of Aldo Manuzio.[12] Given the more standardised form of the printed text, and despite challenges such as frequent ligatures, we included OCR as one of our targets. As a corpus, we selected the *Epistolae diversorum philosophorum, oratorum, rhetorum* (Manutius 1499; ISTC ie00064000, GW 09367. Type 7:114Gr after GW) and the first printed edition of Galen, *Galeni opera omnia graece* (Manutius, Andreas Asolanus 1525. Type 9:84Gr after GW).

We established two primary objectives for this phase: 1) to attain a minimum recognition rate sufficient for aligning text sequences with the corresponding full-text database entries, and 2) to leverage this matching to create a ground truth for further improvement of the OCR algorithm. For optical character recognition, we used *OCR4all*, a software developed by Christian Reul at the Chair of Computer Science at the University of Würzburg under the supervision of Frank Puppe. The latest version of the software can be accessed online.[13]

The text recognition accuracy may vary depending on the specific layout of the input images. Still, it has proven to be generally sufficient for sequence alignment, meaning that it can accurately match a portion of the image with a previously transcribed text.

Sequence alignment is a process that involves searching for similar strings within large sequences using pattern-based techniques. It is widely used in bioinformatics to identify similar strands of DNA, RNA, and proteins, indicating structural, functional or

12 Sicherl 1997; Davies 1999; Staikos 2016; on the digital analysis of Manutius' fonts see Kahl – Kurowsky 2022.

13 https://www.ocr4all.org.

evolutionary differences.[14] This method has also been applied in the humanities to compare text sequences to detect instances of text reuse. Tools such as *Passim* (developed by David Smith)[15] and *TRACER* (developed by Marco Bücheler and colleagues)[16] have been designed specifically for this purpose.

We described the workflow of our *OCR-Sequence Alignment Tool* in a paper presented at the 6th Conference of the Digital Humanities Association in the German-speaking Area (DHd 2019).[17] The software was developed by Markus Bald in his Master's thesis and supervised jointly by the Chair of Computer Science and the Chair of Classics of the University of Würzburg. The pipeline consists of the following steps:

1) *OCR4all* is utilised for the initial processing of images for text recognition. As a representative sample, an early print edition comprising 13 letters of Plato was selected (ISTC ie00064000, see above). The scanned images were segmented into 302 lines, and OCR recognition was implemented with an approximate error rate of 15% (Fig. 6). 793 transcriptions from the *Perseus Digital Library*[18] were taken as comparison texts.

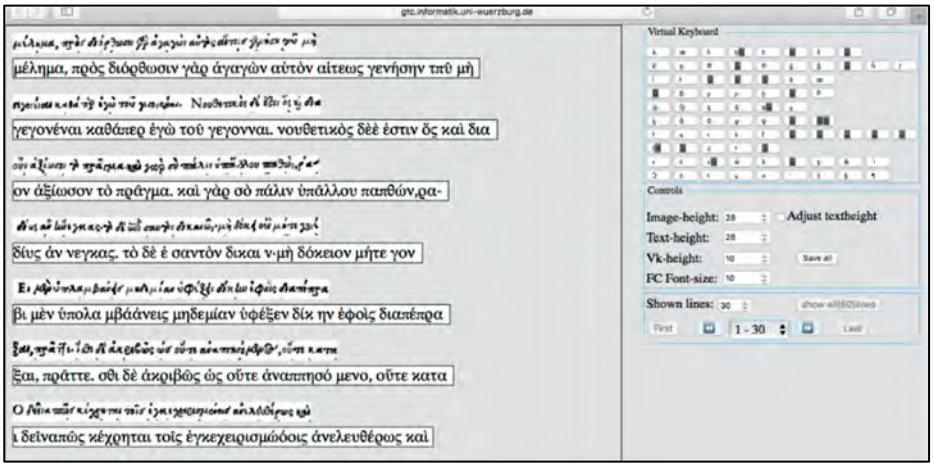


Fig. 6: OCR4all: line segmentation and OCR recognition.

2) Prior to sequence alignment, the lines produced by text recognition are normalised by removing diacritics to increase the likelihood of finding matching equivalents. Based on a comparison of the beginning words of the text, the software preselects the docu-

14 Needleman – Wunsch 1970.
15 https://github.com/dasmiq/passim.
16 https://www.etrap.eu/research/tracer.
17 Bald – Damiani – Essler – Eyselein – Reul – Puppe 2019.
18 https://scaife.perseus.org.

ment with the highest correspondence in the target full-text database. In this way, the search is typically narrowed down to a single work to speed up and facilitate matching.

3) Each line of the input text is then segmented into N-grams of five characters and searched for in the comparison text. Based on clusters of hits in the N-gram search, the software generates potential candidates and assigns them a score based on the number of N-grams found. This process is designed to determine the best candidate in each case. The input and comparison text are matched using the Needleman-Wunsch algorithm (also known as the global alignment technique).[19] The alignment results are then used as input for the correction tool, which displays the original text line alongside the OCR transcription and the best line of comparison text found through automatic alignment (Fig. 7). Any differences between the two versions are highlighted, and users can then choose the correct transcription by selecting individual letters or accepting the comparison text as a whole. The interface includes the printed line from the image, the comparison line selected by alignment, and the OCR-generated line, with coloured markers to define the ground truth; a virtual keyboard with special characters can be activated as needed.

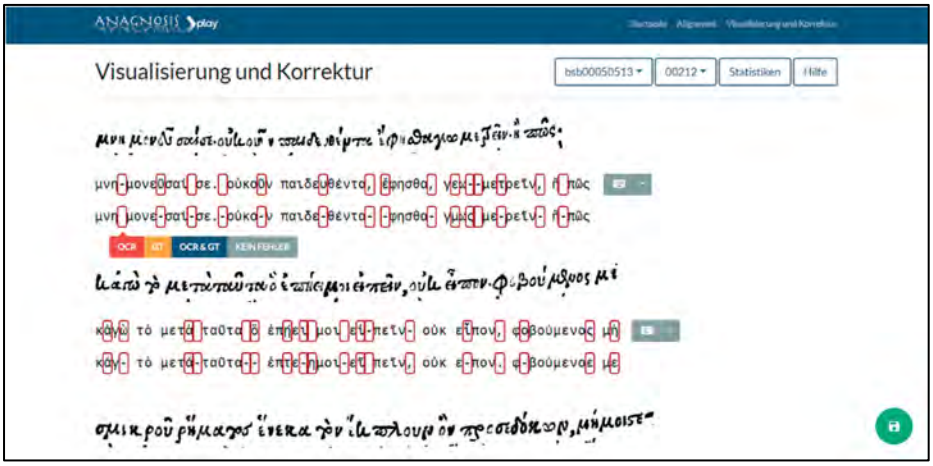


Fig. 7: Correction tool: text-image alignment.

Through automated alignment and manual post-correction, ground truth can be generated more efficiently and used to improve *OCR4all*'s recognition model further.

As a result of the successful processing of early modern prints, the project's scope was expanded in the final phase. The combination of OCR and sequence alignment was applied to produce ground truth for a Greek manuscript from the 15th century (BML,

19 See above, n. 14.

Laur. Plut. 75.7) (Fig. 8), which had never been fully transcribed before. This manuscript was chosen because the shape of the letters does not differ significantly from that of the Aldine editions. This initial attempt provides reason to believe that the systematic and integrated application of both procedures (text recognition and alignment) can significantly speed up the production of ground truth compared to manual processing, which can, in turn, enhance the text recognition of previously untranscribed passages.

Laur. Plut. 75.7 (15th cent.) fol. 1r

1. Laur. Plut. 75.7

Ἀρχόμενοι περὶ τῶν δακέντων ζώων γράφειν τὴν ἀρζὴν ἀπὸ τῶν ἀνθρωποδήκτων ποιησώμεθα, ὡς Ἀρχιγένης φησί, τοὺς μὲν οὖν ἀνθρωποδήκτους πρχρίας αἰσιετον τόπον, μαρα ρίζαν λειὸς μετὰ μέλκατάπλασετὸ ἕλκος μέχρις ἂν γνηται καθαρὸν ταρν τοξδυναμένοις ἀναπληροῦν. οον στὶ τοὐμέλι τερεβιν στέαρ βοστέαρχήνειον· μυελὸς ἐλάφειος· ἰδὲ μὴ παρεη μόχιος τ τήξας χρῶ τδὲ κο ευθὺς ξρχῆς γνοτομελικρα κλύσας ἐπιμελῶς σπόγον κνὸν ἀφρικανὸν μετὰ μέλτ ἐρθοῦ ἀνὰ λαβῶν ἐπιτίθει κἐπίδησον ἐπιδέσει τῃ ἐπικόλπων ρηθησομένη καὶ ἐπίλυε διὰ τρίτης ἢ σμύρναν τερεβινθίνί ἀναλαβν, ἐπιτίθει καὶ ἐπίδησον ἐπιδέσει τῃ ὡς μάλαγμα ἢ χαλκὸν κεκαυμένον λεάνας καὶ

OCR-generated text (OCR4all)

Fig. 8: Original manuscript and OCR transcription.

Efforts were undertaken to integrate the alignment tool into *OCR4all*. In order to ensure maximal compatibility, it will be necessary to adapt the alignment tool to *OCR4all*'s standards in the future. As of now, the alignment of the OCR transcription with comparison texts must be performed separately.

4 Conclusions

The *Anagnosis* project made significant progress over the course of its two phases, resulting in the development of prototypes for valuable tools. It must be acknowledged that the methods presented here may not align with the current state of the art, as the project was concluded in 2019. Nevertheless, the research ideas driving this effort have the capacity to yield valuable insights for future studies and are certainly worth pursuing further. While *Anagnosis* may not have reached its full potential, it has laid the foundation for continued exploration and innovation in the fields of digital papyrology and palaeography.

Bibliography

Ast, R. – Essler, H. (2018), Anagnosis*, Herculaneum, and the* Digital Corpus of Literary Papyri, in *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. by N. Reggiani, Berlin – Boston, 63–73.

Bald, M. – Damiani, V. – Essler, H. – Eyselein, B. – Reul, Ch. – Puppe, F. (2019), *Korrektur von fehlerhaften OCR Ergebnissen durch automatisches Alignment mit Texten eines Korpus*, in *DHd 2019. Digital Humanities: multimedial & multimodal. Konferenzabstracts*, Frankfurt – Mainz, 309–12.

Davies, M. (1999), *Aldus Manutius. Printer and Publisher of Renaissance Venice*, Tempe (AZ).

Kahl, H. – Kurowsky, S. (2022), *Experiment nach Burnhill zur Dimensionierung früher Typographie bei Aldus Manutius*, Digital Classics Online 8, 1–22.

Kiessling, B. – Stökl Ben Ezra, D. – Ast, R. – Essler, H. (2019), *Aligning Extant Transcriptions of Documentary and Literary Papyri with Their Glyphs*, 29th International Congress of Papyrology, Lecce. https://d-scribes.philhist.unibas.ch/en/events-1/neo-paleography-conference/poster-session.

Lowe, D.G. (1999), *Object Recognition from Local Scale-Invariant Features*, in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, II, 1150–7. [https://doi.org/10.1109/ICCV.1999.790410]

Lowe, D.G (2004), *Distinctive Image Features from Scale-Invariant Keypoints*, International Journal of Computer Vision 60, 91–110. [https://doi.org/10.1023/B:VISI.0000029664.99615.94]

Needleman, S.B. – Wunsch, C.D. (1970), *A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins*, Journal of Molecular Biology 48, 443–53.

Ojala, T. – Pietikäinen, M. – Harwood, D. (1994), *Performance Evaluation of Texture Measures with Classification Based on Kullback Discrimination of Distributions*, in *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, I, 582–5.

Ojala, T. – Pietikäinen, M. – Harwood, D. (1996), *A Comparative Study of Texture Measures with Classification Based on Feature Distributions*, Pattern Recognition 29, 51–9.

Sicherl, M. (1997), *Griechische Erstausgaben des Aldus Manutius. Druckvorlagen, Stellenwert, kultureller Hintergrund*, Paderborn – Munich – Vienna – Zürich.

Staikos, K. (2016), *The Greek Editions of Aldus Manutius and his Greek Collaborators (c. 1494-1515)*, Athens.

Isabelle Marthot-Santaniello — Olga Serbaeva

Digital Palaeography of *Iliad* Papyri, D-scribes Project and the Research Environment for Ancient Documents (READ) Platform

1 Introduction

1.1 Goal of D-scribes

The project "Reuniting fragments, identifying scribes and characterising scripts: The Digital Palaeography of Greek and Coptic Papyri (D-scribes)" was funded by the Swiss National Science Foundation (SNSF) and took place in Basel between September 2018 and May 2023.[1] Conceived as a pilot-project, it aimed at investigating how new computational methods could assist papyrologists in identifying handwriting similarities across various fragments and what online interfaces could be the most useful for this task. The project goal and infrastructure are described in Marthot-Santaniello 2021, along with the state of online resources regarding the palaeography of Greek papyri when the project started.

D-scribes focused on three complementary case studies: the papyri bearing Homer's *Iliad*, the archive of Dioscorus of Aphrodito and the archive of Papas. Papas' archive was investigated in collaboration with the research project "Edfou au VIIe siècle" (Edfu in the 7th century) led by Anne Boudhors and Alain Delattre in the French Institute for Oriental Archaeology (IFAO).[2] Dioscorus' archive provided material to evaluate performances of state-of-the-art computational methods in Writer Identification.[3] Except, at the end, an illustration of method transfer to Dioscorus material, the present contribution will focus on the case study devoted to the *Iliad* papyri.

1 Ambizione Grant PZ00P1-174149.

2 https://www.ifao.egnet.net/recherche/operations/op19243. All hyperlinks last accessed on 9.7.2024.

3 A series of articles have been published following the release of a dataset tailored for Writer Identification, see Mohammed – Marthot-Santaniello – Märgner 2019, Christlein – Marthot-Santa-niello – Mayr *et al.* 2022, and most recently Cilia – D'Alessandro – De Stefano *et al.* 2024. For more details on the work done by D-scribes project on this topic, see https://d-scribes.philhist.unibas.ch/ en/case-studies/dioscorus/.

https://doi.org/10.1515/9783111070162-019

1.2 The *Iliad* case study

Papyri bearing Homer's *Iliad* have been chosen as case-study because the *Iliad* is the ancient literary work by far the most attested on papyri with almost 1,500 witnesses[4] spanning over the papyrological millennium. This corpus is thus representative of the chronological evolution of Greek scripts (especially bookhands), but also of their various utilizations, from school exercises to expensive calligraphic books. The goal of the case-study was to investigate how to computationally cluster papyri according to the similarity of their script, the most similar being fragments belonging to the same original manuscript, then fragments from the same writer and last manuscripts belonging to the same style.

The project team started by collecting available data: first, metadata of all the *Iliad* papyri from Trismegistos (Authorwork 511) were exported as a csv file and a manual collection of images was performed from online catalogues, with a priority given to those displaying open licences. After a few months, this work yielded more than 500 images.

1.3 The first experiments

Experiments started first on binarization (understood as segmentation, i.e. separation of the writing from the background), a usual preprocessing step in Computational Analysis of Handwriting. It led to the contribution of D-scribes to the Competition on Document Image Binarization (DIBCO 2019).[5] Papyri, as expected, proved to be challenging to binarize because of their complex background but also of their various degradations and digitization methods. A second experiment in Summer 2020 consisted in evaluating the performances of Handwritten Text Recognition (HTR) using the Transkribus platform in the hope that it could speed up the process of linking text and image. Here again, papyri confirm their challenging nature.[6] Due to limited time and man-power, another approach was chosen in Spring 2021: to be sure to ask the machine to compare what is comparable (and not random patches of various content size), choice was made to proceed to a manual annotation of the *Iliad* papyri at the character level, in order to be able to evaluate similarities between, for instance, all the alphas of the *Iliad* corpus.[7] To efficiently produce these annotations, the platform named the Research Environment for Ancient Documents (READ) was chosen after discussing with Olga Serbaeva, who joined D-scribes project during this period, and Stephen White, one of the main developers of READ.

4 There are 1448 direct attestations in https://www.trismegistos.org/authorwork/511.

5 See Pratikakis – Zagoris – Karagiannis *et al.* 2019.

6 See Marthot-Santaniello – Hodel forthcoming.

7 This approach was strongly influenced by the work of Peter Stokes with Archetype, see https://kdl.kcl.ac.uk/projects/archetype.

2 Annotating papyri in the Research Environment for Ancient Documents (READ) platform

2.1 The choice of READ for D-scribe project

The Research Environment for Ancient Documents (READ)[8] is an Open Source web platform conceived for online editions of manuscripts. It allows linking images with (possibly alternate) text transcriptions, generating a glossary, adding translations and commentaries. The website https://readworkbench.org presents READ features and workflows.

How READ works has been described in Serbaeva – White 2021 and in an online tutorial.[9] Several projects are using READ, among which gandhari.org, the Buddhist Manuscripts from Gandhāra Project[10] and READ Latin.[11] READ can potentially work with any kind of scripts: alphabetic (Greek, Latin), syllabic (Indic scripts), logographic (Mayan).

Created originally for the analysis of early Indic scripts, READ is a perfect tool to analyse single script and single language documents, and it might be particularly helpful with only partly deciphered scripts. It is made for the use-case when a single researcher/team either edits for the first time or updates the reading of an earlier edition of one single historical document at a time.

Since the scope of D-scribes project was not to produce complete online editions of the *Iliad* papyri but to focus on their palaeographical aspects, it was decided from the beginning to use only the features of READ that are relevant to digital palaeography, which include the stable link between the html surface of the historical document and the edition at the character level.

Let us briefly describe the general workflow of the D-scribes project and the place of READ within it. Figure 1 shows that READ was the main Virtual Research Environment for the project.

2.2 The selection of the sources: availability of images and transcriptions

The project used various kinds of sources, the principal being the digital images of the papyri obtained from various institutions. When the IIIF manifest was available, the URL of the image was copied into READ, and the image thus was available within the READ framework without the need of reduplication. When IIIF was not available, the project members collected digital images in various ranges of quality and resolution.

8 https://github.com/readsoftware/read.
9 https://prezi.com/view/f0UoGBtBCWbL4TKi2bVQ.
10 https://www.en.gandhara.indologie.uni-muenchen.de/index.html.
11 https://pric.unive.it/projects/read-latin/home.

These, after receiving necessary permissions and checking the reuse licences, were uploaded to the UNIBAS IIIF server of the project or directly to READ as JPGs.

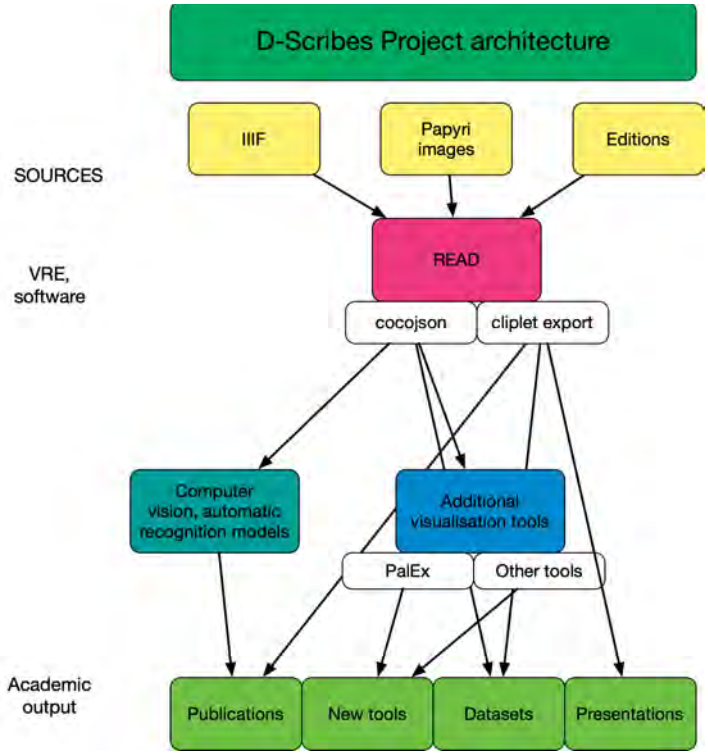


Fig. 1: D-Scribes project architecture.

The project, as a principle, did not work with the images that institutions refuse to publish under open licence or put such conditions concerning, for example, quality, that make the reuse of images meaningless for the project. In those unfortunate cases, the images that could have received more scholarly attention and get rich metadata from the project, just continued to sit in a box, not affected by any new palaeography development.

Compared to documentary papyri that are almost exhaustively available on papyri.info, only a small portion of the literary papyri are encoded in the Digital Corpus of Literary Papyri (DCLP).[12] High quality, open-source images with encoded transcrip-

12 A search on papyri.info for "Ilias" in the "Work" field and "true" in the "Has Transcription" field yields 180 results (last checked on 9 July 2024).

tions were selected in priority. In absence of transcription in DCLP, *Iliad* text was imported from Perseus Digital Library which chose as reference Monro – Allen 1920.[13]

2.3 READ workflow: linking text and image

The sources (i.e. digital images) were included in READ with minimal metadata, often consisting only in the name of the image and the associated TM number.[14] The reason for this minimalism is the fact that READ, in its present version, does not allow the users to access with ease the Postgres database behind or to modify the data structure.

Having connected the image via IIIF URL, or, in some cases, having simply uploaded a JPG to READ (preferred format for READ), the team adds the transcription (from papyri.info or Perseus Digital Library), which has been previously diplomatised, i.e. the signs that are not effectively present on the papyri surface, such as lacuna restauration, modern punctuation or diacritical signs, are removed. The text was converted into upper case because, as usual for papyri, there was no opposition between small and capital letters in the dataset. A small script was created that replaced the letters with accents, spirits and subscript iotas by their root-forms and all instances of sigmas by lunar sigmas. Each line was named/numbered according to the book and verse of the *IIiad* it contained. The diplomatization was finished in READ during the linking process for two main reasons. First, a few papyri presented variant readings (additional letters, different words, plus verses) compared to the vulgate, and these variants had to be added manually. In other cases, some letters that the editors read on papyri were so damaged that they could be misleading for computer vision scripts, and the team chose to mark those as absent/unreadable in the transcription.

READ could not support fully Leiden editorial conventions (it adopts Turfan system[15]) and thus required to accept some noteworthy differences. Round brackets ( ) surround the signs given in the transcription, but not present on the surface (instead of square brackets in Leiden system). Square brackets [ ] are used like underdots in the Leiden system to mark the text that might have a different reading because of the damage. The underdots used on papyri.info to mark uncertainty were thus replaced by [ ] in READ. Because the D-scribes project works with computer vision models, it is important to separate the lines, and thus, the interlinear insertions in Greek that could have been marked with < > received instead their own line numbers. In READ, each used character (letter of the alphabet or symbol) had to be included directly into the code on many different levels, allowing recognition and control of what can be validated or not. This

13 We would like to thank Monica Berti who guided us through the available resources to https://github.com/PerseusDL/canonical-greekLit/blob/master/data/tlg0012/tlg001/tlg0012.tlg001.perseus-grc2.xml

14 Without additional specification, TM numbers are understood as the stable and citable identifiers of texts provided on https://www.trismegistos.org/tm.

15 A basic set of convention is available here: https://gandhari.org/dictionary?section=preface.

approach is motivated by the concern of keeping the data clean, but does not allow the scholar to spontaneously add to the edition the signs that are not in the grapheme table of the code. Since it was im-possible to know in advance all the symbols that will be encountered on papyri, and that many of them are devoid of Unicode characters, the team decided to use ? to mark symbols, conjunct letters, punctuation marks, and other special signs and drawings that do not belong to the 24 letters of the Greek alphabet. These were exported but not separated into subclasses for the present project.

Having uploaded the image and the transcription, the team members drew bounding boxes (bbox) around each letter/symbol on the html surface of the papyri following the order of the lines and linked them to their corresponding letter in the transcription. It is a manual process, and one needs about 10 minutes for 500 signs. In this process, each annotation of bbox type received a unique id number in the dedicated part of the Postgres database.

The linking was verified with the help of the READ palaeography report, which allows to eliminate the intruders and correct any mistakes in linking or in the edition, see Figure 2. There is a possibility to export cliplets, understood as small PNG images of individual characters. This option allowed the generation of frozen datasets for Machine Learning approaches but was also useful for presentations and visuals for publications.

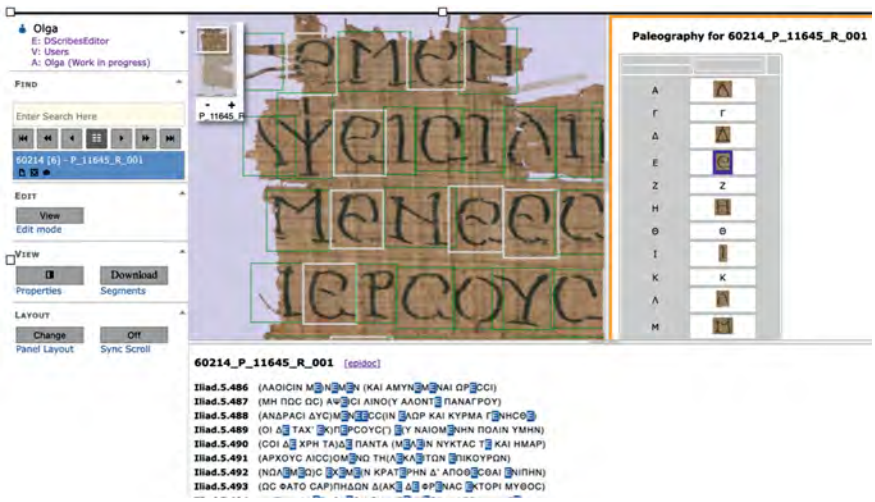


Fig. 2: TM 60214 in READ backend showing the link between the surface of the image, the transcription and the palaeography report in which epsilon is selected.

2.4 The cocojson

In order to reuse READ data in a better structured way, a cocojson export option was created by Stephen White and Selaudin Agolli (student assistant in D-scribes project).

The cocojson export file comprises the following blocks: “annotations”, “categories”, “database”, “images”, “licenses”, “texts”.

The “annotations”, i.e. segment id linked to their bbox coordinates for each annotated letter or symbol, are of the type:


```
 {
       "area": 5135,
       "bbox": [
       2259,
       417,
       79,
       65
       ],
        "category_id": 23,
        "id": 1,
        "image_id": 1,
        "iscrowd": 1,
        "seg_id": 84,
        "tags": {
        "BaseType": [
        "bt2"
        ],
        "FootMarkType": [
        "ft1"
        ]
},
```


Let us briefly explain the most relevant parts: “bbox” is the annotation that marks the 0 point of the x,y coordinate system (first two numbers marking the upper left corner of the annotation box). The other two numbers, again in order of x,y mark the difference from the 0 point respectively to the right and down. The values are in pixels. Tags here stand for annotations of the preservation quality and letter shapes, and they will be described in details below.

The “category_id” above refers to the Greek letter or symbol from a controlled list of symbols hard coded in READ under “categories”.[16] For example:


```
{
    "id": 8,
    "name": "α",
    "supercategory": "Greek"
},
```


16 As of July 2024, Stephen White, the principal READ developer, has informed us that there is a plan to provide a tool that would allow scholars to create/update grapheme tables from READ with ease. So far, the creation or modification of these tables had to be done by modifying READ code directly.

While “image_id” refers to the image encoded as follows:


```
        {
            "bln_id": 1,
            "date_captured": null,
            "file_name": "H1 PCI MSS A101 XIII.jpg",
            "height": 2162,
            "id": 1,
            "img_url":          "https://app.d-scribes.philhist.unibas.ch/im-
    ages/homer2/ txt1/P.Corn.Inv.MSS.A.101.XIII.jpg",
            "licence": 1,
            "width": 2524
        }.
```


The image has licence, listed under “licences”, and the whole structure is finally linked to “texts”, i.e. editions:


```
        {
            "tm": "60701",
            "txt_ckn": "60701 [6]",
            "txt_id": 1,
            "txt_image_ids": [
                1
            ],
            "txt_ref": "PCI MSS A101 XIII\n",
            "txt_title": "PCI MSS A101 XIII / P_Oxy_III_549"
        }.
```


This data structure can be immediately reused for training computer vision models as well as it can be shared with and reused by other scholars.

2.5 Tagging

A key feature of READ is the possibility to assign tags to the individual letters. Three kinds of palaeographical tags were used: BT, FT and VT. Their names come originally from a syllabic language project (Gandhari), i.e. Base type (BT), Footmark type (FT), and Vowel type (VT) but were used differently in the context of Greek, as exposed in what follows.

2.5.1 BT-tagging

For the *Iliad* project, it was decided from the beginning to annotate all the letters, even the very damaged ones, with the exception of a few hardly visible ones that editors had underdotted and that the team decided to treat as lost, as mentioned above. It was also acknowledged as relevant to be able to differentiate various states of preservation.

Therefore, BTs would define the quality of a given cliplet containing an individual letter. BT1 was assigned to the highest quality, when the letter had clear, complete or close to complete shape.

BT2 was assigned when a letter had some damage that might complicate computer analysis, i.e. when parts of neighbouring letters would be present or when the pattern of the letter is incomplete due to holes, breaks or erasure of ink. The criterion for assigning the BT2 tag is that, even if the letter is damaged, its shape alone allows the identification to only one out of the 24 Greek letters without external knowledge (use of the context, transcription).

BT3 was assigned when what remains of the letters could be variously interpreted and the transcription is required to only assign one letter type (for instance a vertical stroke that can either be a iota or part of beta, gamma, kappa, etc). It is, however, different from BT4 which was used when the damage is such that it could lead to a misreading, for instance, an alpha with its middle bar erased, or a delta with its horizontal stroke damaged would look exactly like a lambda. Examples of the BT taggings of epsilon are provided in Figure 3.

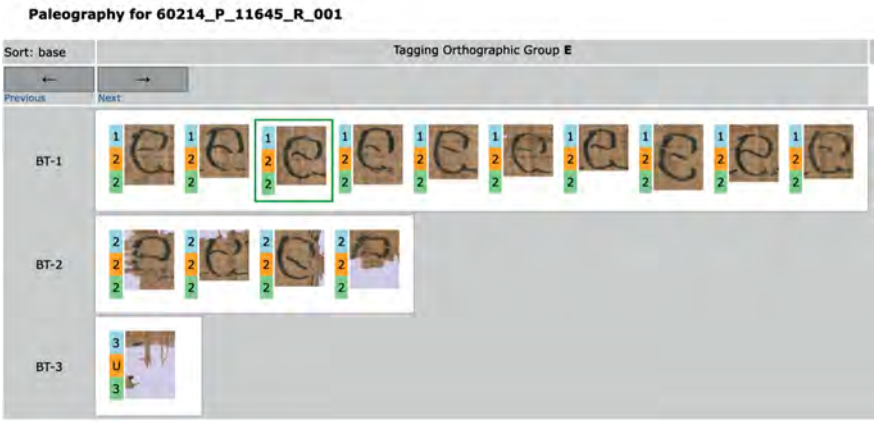


Fig. 3: BT-tagging of the epsilon from a papyrus belonging to TM 60214.

2.5.2 FT-tagging

The second round of tagging had the aim to link together the letters of similar shape (understood as structure) by assigning them the same FT-type. The typology for this experiment was rather naive, created locally by putting similar letters exported from READ as PNGs into the boxes (files), see Figure 4.

The aim of this tagging was to test if it would be possible to cluster the documents without much resources by assigning a dominant FT-type to each letter class (for ex-

ample alpha) found in papyri. We would then have a portrait of each fragment in terms of the dominant types of each letter (Fig. 5).

A	FT1	FT2	FT3	FT4	FT5	FT6	FT7
	Like "d", one stroke, roundish belly	2 strokes, angular belly, there is space on the upper line before connecting the belly	3 strokes, like A	not having the space before belly on the upper line	with loop on the left leg, like capital	like FT2, 2 strokes, but with round belly	2 strokes, angular belly, the belly is written in one line, slanted
n. of strokes	1	1-2	3	1-2	2	2	2
							

Fig. 4: Original naive classification of alpha reflected in the FT-tagging.

	α	υ	χ	ζ	ξ	γ	ι	δ	η	κ	λ
175242	6	no υ	no χ	no ζ	no ξ	no γ	2	no δ	1	2	no λ
60214	6	3	1	no ζ	no ξ	no γ	2	1	1	no κ	1
60215	3	2	1	2	4	1	2	1	1	2	1

Fig. 5: FT-based dominant letter types for each TM and each letter (small selection). In red are the letters not found on papyri. Numbers stand for FT types.

Although this method gave some promising results on visualisation, it was not sufficiently fine-grained. Any missing letter class could potentially undermine the clustering results. Besides, many papyri had two and more types for the same letter class. Here is a sample visualisation of the FT-tags of 23 randomly chosen items from the *Iliad* dataset (Fig. 6), where the similarity of the shape of skyline would be a proof of the similarity of two or more TMs.

Some of those differences could be explained by the letter position (initial or last of the line, distortion due to semi-cursivity), but in many cases, this fluidity of shape within one fragment was clearly the expression of the scribe's freedom, and a better method, along with a different typology, was needed.

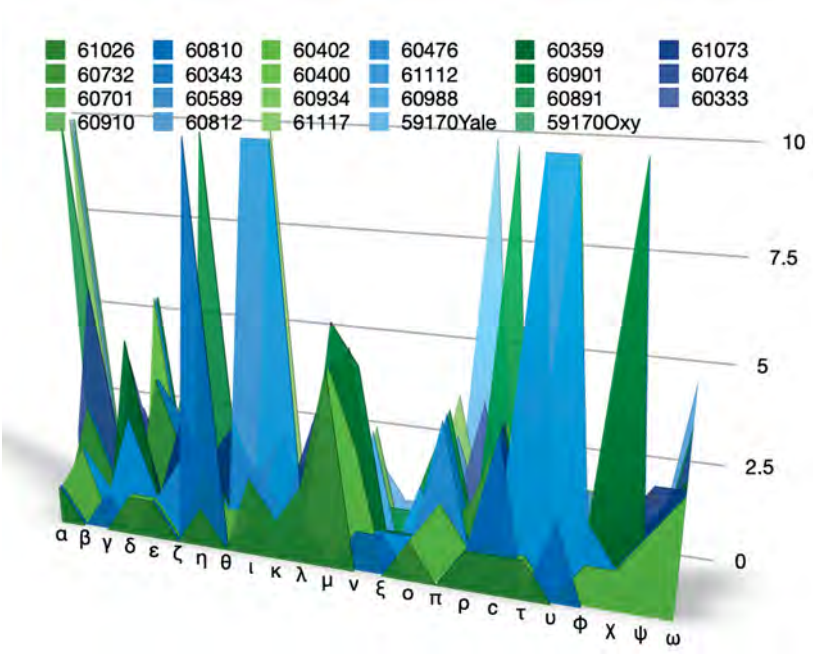


Fig. 6: The skylines of the randomly selected TMs (alpha to omega order) based on the FT-tags (0 to 6). When an item has multiple types of the same letter, value 10 was assigned to that letter.

2.5.3 VT-tagging

A new typology came into being, it was based on allographs, defined as letter shapes resulting from different ductus understood as the number, direction and order of the strokes. In the previous experiment it was clear that some letters might look similar, but they are written with a different number of strokes, or, on the contrary, that some letters might look very different, but are a gradual, logical evolution from one another, like a calligraphic beta of 4 strokes that looks like Latin B can gradually evolve into a U shape.

Thus, the team came up with an updated hierarchical structure that was used for the VT-tagging (tested, but without covering the whole *Iliad* dataset present in READ). Here is an example of alpha VT-tagging hierarchical tree with the sample cliplet export from READ (Fig. 7):

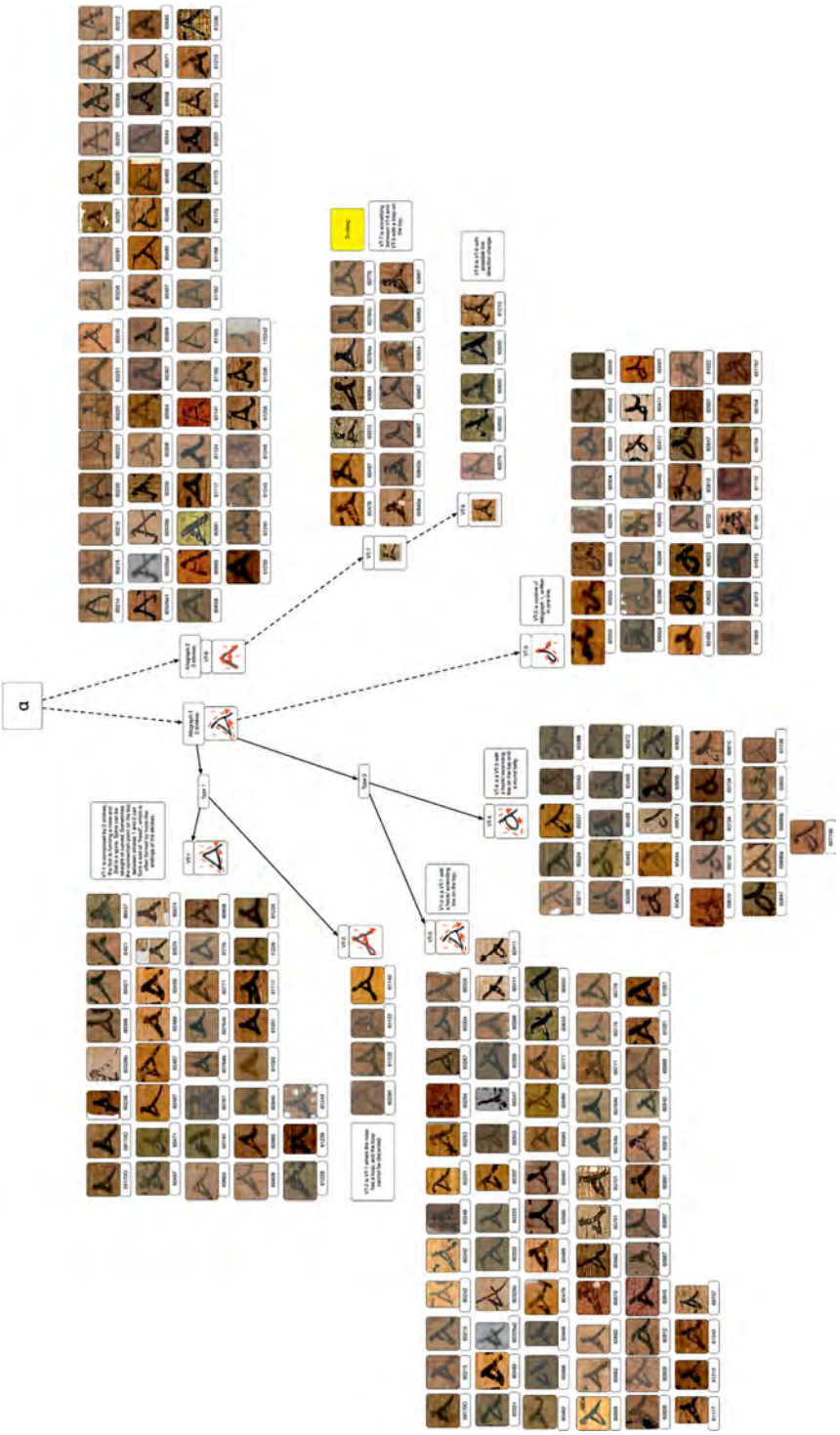


Fig. 7: Hierarchical representation of alpha VT-tagging by allographs.

A few statistical tests on the output of the tagging were done and the visualisation of the trees effectively showed that this simple method allowed to link together the images that belonged to the same TM. However, the weak side of this method was the fact that if a fragment had many missing letters, there was not enough ground to be able to link it to another fragment. This method tried to reduce complexity: one would have a portrait of the papyri defined in terms of 24 parameters (24 Greek letters). The problem of missing letters was mitigated by assigning the value that is most likely to appear in the closest items where it is present.

With this method we could get some distant view results, such as the fact that the items of Biblical majuscule style were clustered closely together (in red) (Fig. 8).

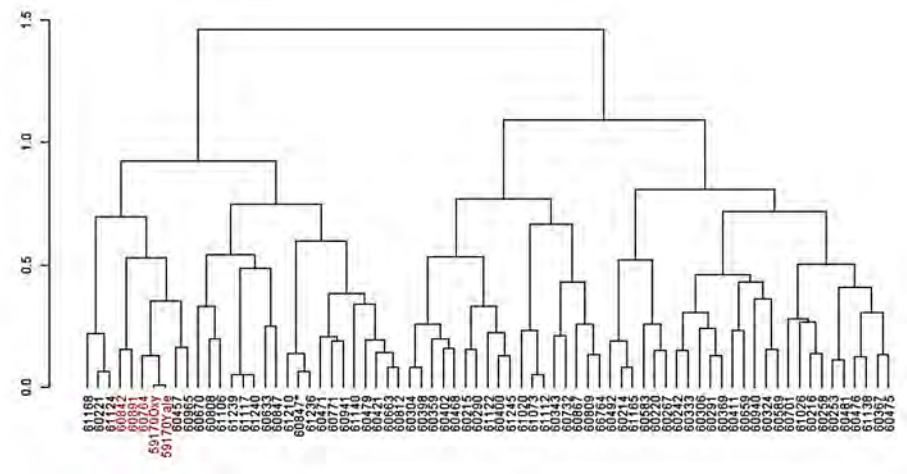


Fig. 8: Clusters of VT-tagging of 3 letters (A,E,M), of 72 items belonging to the *Iliad* dataset, done to evaluate the results of computer vision analysis presented in Marthot-Santaniello – Tu Vu – Serbaeva – Beurton-Aimar 2023. In red are the papyri written in Biblical Majuscule.

This method, which can be run on a single computer and requires very little resources, might still be used for the preliminary exploration of the datasets.

3 The relevance of READ annotations for palaeography and computational approaches

3.1 PalEx, a new visualisation tool for READ data

As soon as the accent moved from working on a single papyrus to finding similarities and differences in hundreds of them, D-scribes team needed a new visualisation tool that would be faster than READ, more reactive, and that would allow navigating between letters and items instantly. Such a tool, named PalEx (for **Pal**aeographic **Ex**plorator), was created by Selaudin Agolli under the supervision of Stephen White.[17]

PalEx digests the READ cocojson and adds new features to the Palaeography report visualisation compared to READ. For instance, it allows the user to see all bboxes simultaneously, with different letters marked by different colours, as below (Fig. 9):

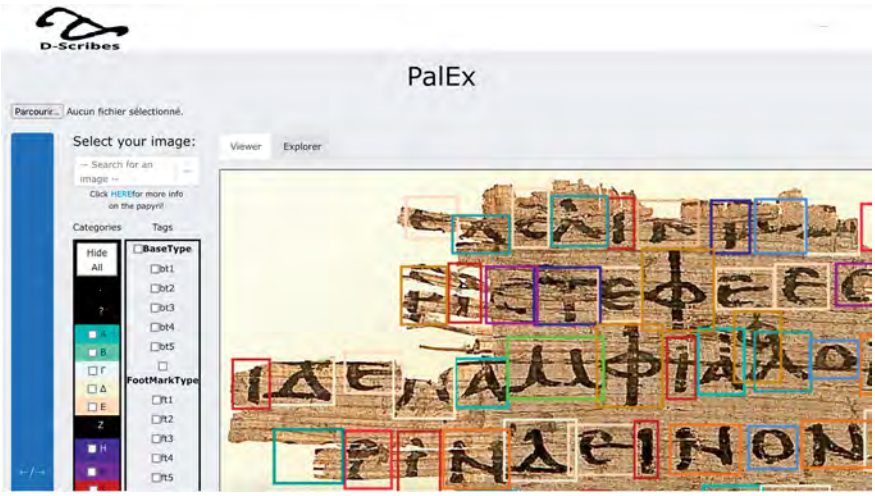


Fig. 9: A landing page view in PalEx, with the first document TM 60701.

PalEx offers an option of multiselect, and the scholars can freely combine the letters and the tags in selection. Below is an example of selection of A and Φ (Fig. 10):

17 PalEx is available on https://showcase.d-scribes.philhist.unibas.ch/palex/coco/1 (frozen dataset containing *Iliad*) and on the project webpage, where it is possible to explore in PalEx any READ-like cocojson: https://showcase.d-scribes.philhist.unibas.ch/palex.

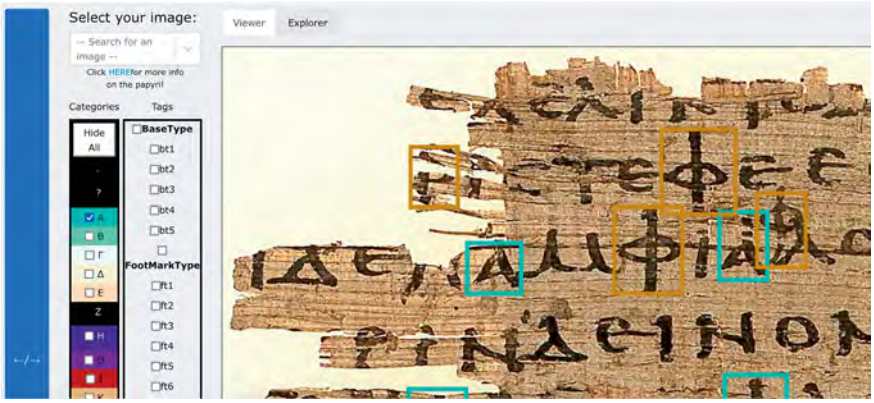


Fig. 10: TM 60701 with multiple selection visualised.

PalEx naturally incorporates the best parts of READ palaeography report, allowing to explore various tags (Fig. 11):

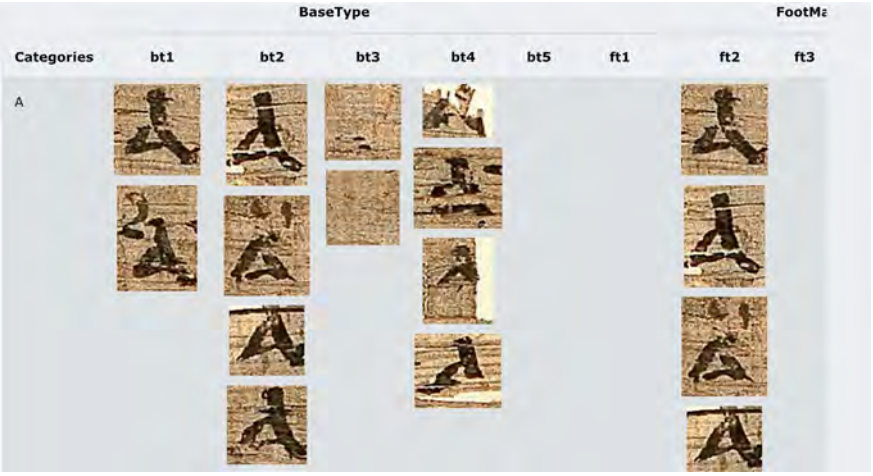


Fig. 11: Alphas from TM 60701 with their BT- and FT-tags in the Explorer mode of PalEx.

When the user clicks on any cliplet, the full image centred on that cliplet opens, providing thus the immediate context. This interface has proven to be extremely useful for palaeographical comparisons of *Iliad* papyri suspected to be either from the same manuscript or at least the same writer.[18]

18 Several such cases have been spotted during D-scribes project and an article is in preparation.

3.2 Towards an automatic annotation of papyri: the competition on Detection and Recognition of Greek Letters on Papyri (DRoGLoP)

Following the procedures explained above, D-scribes team encoded 150 TMs containing part of the *Iliad* in READ at the character level during one year.[19] The process was too time consuming to be pursued and a search for automatic solutions started. The careful manual annotation work linking image and transcription at the character level had a great potential to train computer vision models aiming at automatic detection and recognition, as it provided the necessary ground truth. A competition was organised during the 17th International Conference on Document Analysis and Recognition (ICDAR) based on 187 images belonging to 139 TMs along with the cocojson and an explanation on the use of BT tag.[20] The promising results obtained by the winning teams[21] allows hoping that, in a near future, only a small manual curation will be required on the output of automatic detection and recognition for at least literary and non-cursive scripts.

3.3 Stylistic similarities among letter shapes: two Machine Learning experiments

Thanks to the possibility offered by READ to export selections of cliplets, an experiment was led in collaboration with Manh Tu Vu and Marie Beurton-Aimar (LaBRI, Université Bordeaux) and published in Marthot-Santaniello – Tu Vu – Serbaeva – Beurton-Aimar 2023. Among the *Iliad* papyri annotated in READ, 72 TMs were selected because they contained two or more high quality specimens (BT1) of alphas, epsilons and mus, forming the AlphEpMu dataset of more than 5,000 cliplets.[22] After training a neural network called SimSiam[23] by defining as similar letters coming from the same papyrus, similarity estimation results were obtained that could be visualised as graphs (Fig. 12). Several groups corresponded to traditional palaeographical categories. These encouraging results should be confirmed in the future by experiments on larger datasets with more numerous letters than only alpha, epsilon and mu.

19 Prezi presentation on *Iliad*: https://prezi.com/view/cecwVpsKBiPB5u80ecZp.

20 Seuret – Marthot-Santaniello – White *et al*. 2023.

21 See for instance Vu – Beurton-Aimar 2023.

22 AlphEpMu dataset available at https://d-scribes.philhist.unibas.ch/en/case-studies/iliad-208/alphepmu-dataset/.

23 Original code: https://github.com/Papyrus-Analysis/writer_verification_network/tree/simsiam-paper.

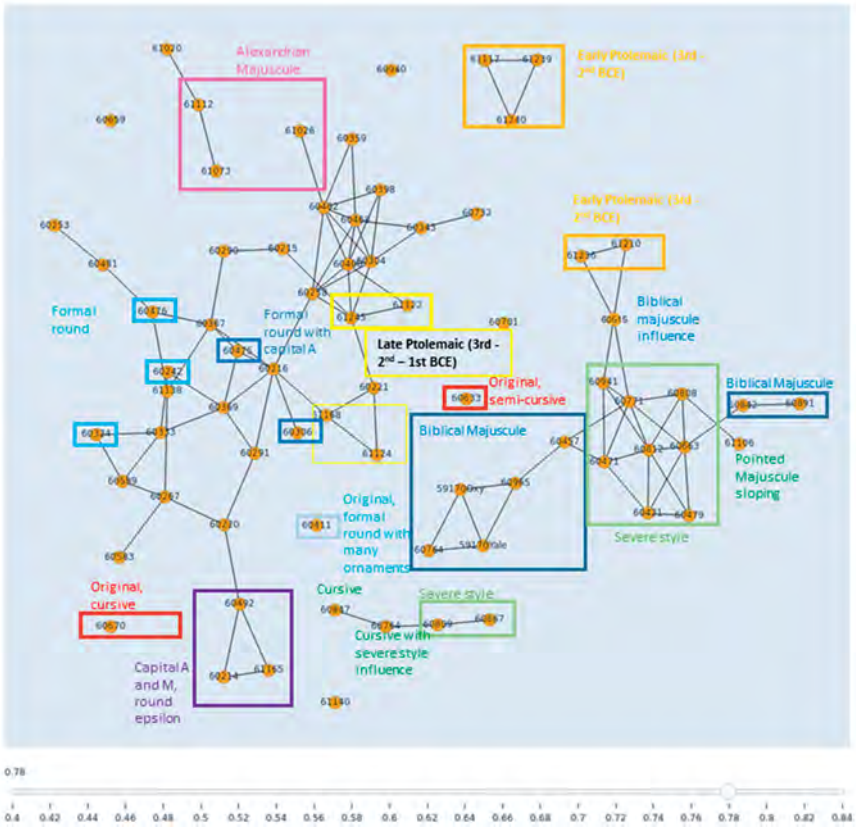


Fig. 12: Visualization of the similarity estimation results (threshold set at 0.78) among 72 *Iliad* papyri based on similarity scores of alphas, epsilons and mus, see Marthot-Santaniello – Tu Vu – Serbaeva – Beurton-Aimar 2023.

The code created in the frame of this collaboration with the University of Bordeaux was later reused and upgraded by Giuseppe De Gregorio in a work focusing on a smaller Biblical Majuscule corpus which also benefited from READ export of clip-lets.[24] The goal of this experiment was to evaluate levels of similarities among seven *Iliad* manuscripts written in Biblical Majuscule, combining computer-assisted paleography and computational analysis. It allowed spotting a group of texts that shared strong similarities but also a fragment that most likely did not belong to the manuscript it was assigned to. The results did show considerable difference of shape between the letters written by the

24 See García-Baró – de Gregorio – Serbaeva – Marthot-Santaniello forthcoming. Video of the conference presentation accessible at https://d-scribes.philhist.unibas.ch/en/events-1/papyri-conference/conference-videos/

same scribe. This finer analysis that went into the differences rather than into the similarities, allowed to produce a Gephi visualisation of the clustering, where it was clear that some letters of a manuscript can be closer by their shape to letters of another manuscript than to its own. These two experiments, although still preliminary, validated the relevance of character level annotations and underlined the importance of rigorously determining meaningful thresholds in similarity scores. This approach aims at allowing in the future definitions of "hand", "writer" and "style" that can contain reproducible, metrical values.

3.4 Beyond *Iliad*: READ annotations on Dioscorus archive

In order to see the problem of hand variation within the documents written by a single scribe, another dataset was added to READ at the end of D-scribes project. It included selected documents from the GRK-120 dataset produced in the scope of Writer Identification research on Dioscorus archive.[25] It is a rare case where we do have an archive that is localised in time (documents range from 500 to 650 AD) and space (the village of Aphrodito was situated near present day Sohag). GRK-120 includes documents attributed to 23 scribes and the first Machine Learning experiments based on random patches extracted from images of various definitions yielded fragile results, suggesting thus that manual annotation could significantly improve the performances.

Because of time constraints, it was impossible to annotate each character of those 120 documents, so two small datasets were created in READ. The first included tagging of about three lines of each document (around 200 characters, to include the whole alphabet when possible, or, at least, the largest part of it) in the same way as it was done for *Iliad*. This allowed to create an overview of the alphabet written by each scribe and the variations in the letter shapes. The second dataset consisted only of και, meaning "and" or being a part of words such as δικαιος. Και was chosen because statistically it is the most frequent three-letter combination in Greek and because its shape is in most cases not distorted by what precedes or follows, even in cursive scripts.[26] Και was annotated within a single bounding box including all three characters, because it was important to assess if letter combination would work better for computational Writer identification than single letters.

An additional difficulty of the dataset was the fact that most of the documents were written in cursive scripts. Although READ had proved to work relatively well with cur-

25 https://d-scribes.philhist.unibas.ch/en/case-studies/dioscorus/kairacters.

26 Prezi presentation of the dataset is available here: https://prezi.com/view/jwaI0nWJ4IbfbDcz2Rqa.

sive[27], one would wish to have not a bbox but rather a rhombus (for slanted scripts) or a free polygon[28] that would more closely capture targeted shapes.

In a study led in collaboration with Marco Peer and Robert Sablatnig (Computer Vision Lab, TU Wien), Machine Learning methods confirmed that training a neural network with a few manual annotations of coherent content performed better than with massive random patches of various content.[29] The possibility to automatically detect και in any papyrus image will be investigated in the future.

4 Conclusion

READ as a framework served as the main framework for D-Scribes project for 3 years, from 2021 to 2023, and allowed to produce reliable datasets that became ground truth for machine learning, which should reduce the need of manual annotation in the future (work in progress by Stephen White and Giuseppe de Gregorio). It enabled the possibility of multiple tests and explorations, in the domain of Greek palaeography as well as in Computer vision. READ remains an excellent tool for such use cases as the *Iliad* project.

What is on the wish list is the easily configurable data model, the export-import options,[30] and the connections to other tools preferably via restful API. For instance, one would greatly appreciate a pipe-line from READ to Jupyter notebooks/R code that would allow scholars to reuse the massive amount of data produced and stored in READ Postgres database, providing simple statistical analysis and visual exploration for the scholars.

Pilot experiments led in the scope of D-scribes opened the way to a new project called Egrapsa.[31] As the research aims shifted from the analysis of a single document to the large-scale comparison of multiple documents based on their dates, proven-ance and the socio-cultural background of their writers, it became clear that READ alone would not be enough to store and allow to query that much wealth of data. This changed the architecture of the project and the workflow, from a situation of READ as the centre to READ as one among many building blocks. This is, however, a subject for another article.

27 for example, a test was done on a column of P.Bodmer 1 verso in order to virtually restore the missing parts, see Perrin – Cudilla – Xie *et al.* 2023.

28 This option is already available in READ, but it is very time-consuming. Most Machine Learning approaches require square images as input.

29 Peer – Sablatnig – Serbaeva – Marthot-Santaniello forthcoming.

30 The cocojson export is an important development in that direction.

31 SNSF Starting Grant Project: “EGRAPSA: Retracing the evolutions of handwritings in Graeco-Roman Egypt thanks to digital palaeography" (Basel, June 2023–May 2028)”.

Bibliography

Christlein, V. – Marthot-Santaniello, I. – Mayr, M. – Nicolaou, A. – Seuret, M. (2022), *Writer Retrieval and Writer Identification in Greek Papyri*, in *Intertwining Graphonomics with Human Movements. IGS 2022*, ed. by C. Carmona-Duarte – M. Diaz – M. A. Ferrer – A. Morales, Cham, https://doi.org/10.1007/978-3-031-19745-1_6.

Cilia, N. D. – D'Alessandro, T. – De Stefano, C. – Fontanella, F. – Marthot-Santaniello, I. – Molinara, M. – Scotto Di Freca, A. (2024), *A Novel Writer Identification Approach for Greek Papyri Images*, in *Image Analysis and Processing - ICIAP 2023 Workshops*, ed. by G. L. Foresti – A. Fusiello – E. Hancock, Cham, 422–36, https://doi.org/10.1007/978-3-031-51026-7_36.

García-Baró, P. – de Gregorio, G. – Serbaeva, O. – Marthot-Santaniello, I. (forthcoming), *Biblical Majuscule: Computer Spotted Features and Palaeographer's Perception*, proceedings of the online conference "Perceptions of Writing in Papyri. Crossing Close and Distant Readings", 7–8 December 2023.

Marthot-Santaniello, I. (2021), *D-scribes Project and Beyond: Building a Virtual Research Environment for the Digital Palaeography of Ancient Greek and Coptic Papyri*, ed. by C. Clivaz – G. V. Allen, Classics@ 18, https://classics-at.chs.harvard.edu/classics18-marthot-santaniello.

Marthot-Santaniello, I. – Hodel, T. (forthcoming), *Papyri, Handwritten Text Recognition, and Text Processing. State of the Art and Outlook – Approaches to a Digital Paleography*, Proceedings of the international workshop "In the Name of the Rose. Searching for Unknown, Lost, and Forgotten Ancient Texts", Rome, September 30–October 1, 2021.

Marthot-Santaniello, I. – Tu Vu, M. – Serbaeva, O. – Beurton-Aimar, M. (2023), *Stylistic Similarities in Greek Papyri Based on Letter Shapes: A Deep Learning Approach*, in *Document Analysis and Recognition – ICDAR 2023 Workshops*, ed. by M. Coustaty – A. Fornés, Cham, 307–23, https://doi.org/10.1007/978-3-031-41498-5_22.

Mohammed, H. – Marthot-Santaniello, I. – Märgner, V. (2019), *GRK-Papyri: A Dataset of Greek Handwriting on Papyri for the Task of Writer Identification*, in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, 726–31, https://doi.org/10.1109/ICDAR.2019.00121.

Monro, D. B. – Allen, T. W. (1908–1920), eds., *Homeri Opera*, Third Edition, 5 vols., Oxford.

Peer, M. – Sablatnig, R. – Serbaeva, O. – Marthot-Santaniello, I. (forthcoming), *KaiRacters: Character-level-based Writer Retrieval for Greek Papyri*, preprint https://arxiv.org/pdf/2407.07536.

Perrin, S. – Cudilla, L. – Xie, Y. – Mouchère, H. – Marthot-Santaniello, I. (2023), *Homer Restored: Virtual Reconstruction of Papyrus Bodmer 1*, in *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing (HIP '23)*, New York, 37–42, https://doi.org/10.1145/3604951.3605518.

Pratikakis, I. – Zagoris, K. – Karagiannis, X. – Tsochatzidis, L. – Mondal, T. – Marthot-Santaniello, I. (2019), *ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019)*, in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, 1547–56, https://doi.org/10.1109/ICDAR.2019.00249.

Serbaeva, O. – White, S. (2021), *READ for Solving Manuscript Riddles: A Preliminary Study of the Manuscripts of the 3rd* ṣaṭka *of the* Jayadrathayāmala, in *Document Analysis and Recognition – ICDAR 2021 Workshops*, Lausanne, Switzerland, September 5–10, 2021 Proceedings, Part 2, ed. by E. H. Barney Smith – U. Pal, Cham, 339–48, https://link.springer.com/chapter/10.1007/978-3-030-86159-9_24.

Seuret, M. – Marthot-Santaniello, I. – White, S. – Serbaeva, O. – Agolli, S. – Carrière, G. – Rodriguez-Salas, D. – Christlein, V. (2023), *ICDAR 2023 Competition on Detection and Recognition of Greek Letters on Papyri*, in *Document Analysis and Recognition - ICDAR 2023. ICDAR 2023*, ed. by G. A. Fink – R. Jain – K. Kise – R. Zanibbi, Cham, 498–507, https://doi.org/10.1007/978-3-031-41679-8_29.

Vu, M. – Beurton-Aimar, M. (2023), *PapyTwin Net: A Twin Network for Greek Letters Detection on Ancient Papyri*, in *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing (HIP '23)*, New York, 43–8, https://dl.acm.org/doi/10.1145/3604951.3605522.

Nicole Dalia Cilia — Tiziana D'Alessandro — Claudio De Stefano — Francesco Fontanella

Writer Identification from Handwriting on Greek Papyri

1 Introduction

Over time, man has developed a series of well-defined signs capable of fixing an articulated language to transmit it to succeeding generations, thus giving rise to writing. Thanks to it, it is possible to reconstruct an awareness of past events; thus, it represents our most important historical source. The importance of writing in historical reconstruction has spawned a plethora of disciplines devoted to learning, reconstructing, and interpreting these sources: one of these is papyrology.

Studies in the field of papyrology do not focus only on the transcription of every papyrus fragment, but also on other challenging tasks such as digitization,[1] the document enhancing,[2] the writer identification,[3] the recognition of materials used, the definition of the place and the period in which the transcription was executed. This project aims to build an end-to-end system to support researchers during the writer's identification of ancient Greek papyri.[4] In past years, this challenge was faced by different researchers with many approaches.[5] The following paragraphs show the dataset of papyri used and the three different approaches evaluated.

2 Dataset

The reference dataset considered in this work comprises images representing Greek papyri dating back to about the 6th century AD, which have been selected and cataloged by experts in the field of papyrology. All the documents used are part of the richest archive of the Byzantine period, belonging to Dioscorus of Aphrodito, which collects more than

1 See Jayanthi – Indu – Hasija – Tripathi 2022.

2 See Gupta – Kumar – Gupta – Chaudhury – Joshi 2007; He – Schomaker 2019.

3 See Rehman – Naz – Razzak 2019.

4 The project is conducted at the Department of Computer Engineering, University of Enna "Kore" (N. D. Cilia) and at the Department of Electrical and Information Engineering (DIEI), University of Cassino and Southern Lazio (T. D'Alessandro, C. De Stefano, F. Fontanella).

5 See Nasir – Siddiqi 2021; Nasir – Siddiqi – Moetesum 2021; Christlein – Marthot-Santaniello – Mayr *et al.* 2022; Peer – Sablatnig 2023; Cilia – D'Alessandro – De Stefano *et al.* 2024.

https://doi.org/10.1515/9783111070162-020

700 texts written in cursive Greek.[6] The base of this dataset is the GRK-Papyri,[7] which is used to identify the writers and is composed of 50 images distributed unequally among the 10 available writers. Subsequently, other papyri have been added to this starting set, which partly increased each existing writer's number and introduced new authors.[8] Finally, a number of images equal to 122 was reached for a total number of 23 writers.

2.1 Problems

The papyrological transcription and identification tasks are extremely difficult due to the deteriorated condition of most papyrus fragments. Every papyrus is different as each is affected by a certain level of degradation, showing missing pieces, holes, and bending marks. These problems are due to the natural deterioration of the papyrus material or the exposure to harsh natural conditions. The writing may have become illegible, or the ink may have faded with time. Despite being uniquely different in content and composition, they mostly differ because of the deterioration process.

The papyri available in the reference dataset are located and exposed in different parts of the world, representing important documents and part of the artistic and cultural heritage. This means that some images of these papyri may have been acquired with different tools and resolutions and under different light conditions; also, for some of them, it is possible to notice the reflection of the glass that covers them for preservation purposes. Moreover, images have different sizes and are saved with different numbers of color channels; some of them are greyscale, while others are RGB.

Table 1: Writers and the number of papyri that belongs to them.

Writer	no. of Papyri
Abraamios	21
Amais	1
Andreas	4
Anouphis	1
Apa Rhasios	4
Dios	15
Dioscorus	5
Daueit	1

6 See Fournet 2008.

7 See Mohammed – Marthot-Santaniello – Märgner 2019; https://d-scribes.philhist.unibas.ch/en/grk-papyri. All hyperlinks last accessed on 21.4.2024.

8 See Cilia – De Stefano – Fontanella *et al.* 2021.

Writer	no. of Papyri
Hermauos	5
Ieremias	2
Isak	8
Kollouthos	2
Konstantinos	2
Kyros (1)	9
Kyros (2)	1
Kyros (3)	5
Menas	5
Philotheos	3
Pilatos	10
Psates	2
Theodosios	5
Victor (1)	10
Victor (2)	1

3 Experimental procedure

This project aims to build a support system for the writer's identification of ancient Greek papyri. The system relies on Deep Learning (DL) techniques to automatize the process of writer recognition according to handwriting. Because of the problems mentioned in the previous paragraph, a preprocessing step on the images was necessary, trying to uniform them and highlight the handwriting. Following, three different approaches are described. Each one is composed of image preprocessing and a classification step.

3.1 First Approach

Image Pre-processing The image preprocessing step of the first approach tried to solve more than one problem and made the dataset as uniform as possible. The intention was to apply DL techniques on those images, so the dataset needed homogeneity. Also, every papyrus' background had to be as uniform as possible. First, the background uniformity was obtained by filling the holes with the same color as the papyrus background and turning the image greyscale.

Then, images were resized and rotated, if necessary, to extract rows of different lengths. After this step, two datasets were obtained from the initial one:

– a dataset of extracted rows of 1232 pixels width;
– a dataset of extracted rows of 500 pixels width.

Classification The classification step required a well-balanced dataset among the various classes (writers); this led us to arrange the data in several ways according to the number of writers to recognize and the availability of images for each writer. That is why the classification phase was performed on five arranged datasets from the two sets of images discussed in the previous section. We adopted a K-Fold Cross Validation strategy to improve the system's stability. Four Convolutional Neural Networks (CNNs), pretrained on the public dataset ImageNet,[9] were tested considering the DL techniques of Transfer Learning and Fine Tuning. The chosen CNNs are VGG19,[10] ResNet50,[11] InceptionV3[12] and InceptionResNetV2.[13]

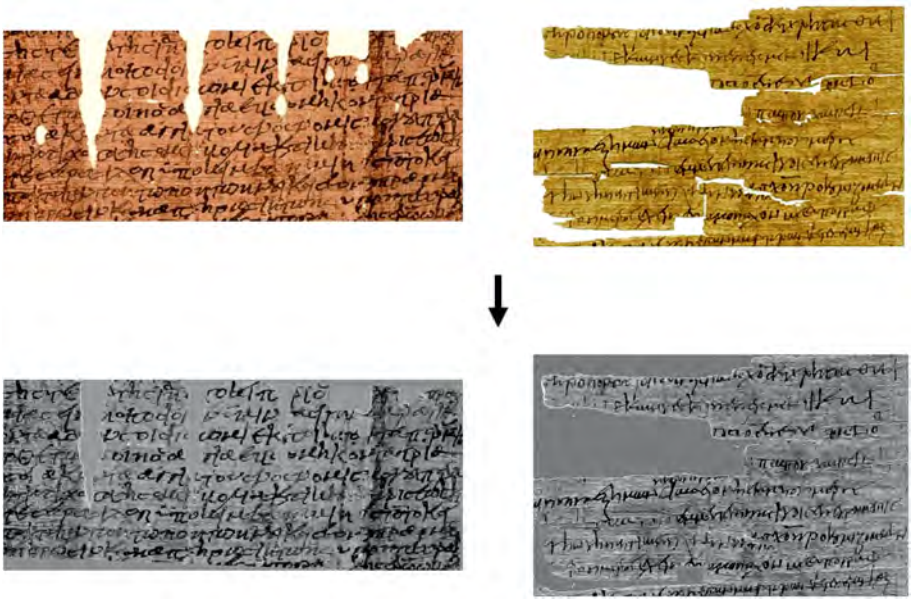


Fig. 1: Processing on original images turned into greyscale with background uniformity.

9 See Deng – Dong – Socher *et al.* 2009.
10 See Simonyan – Zisserman 2015.
11 See He – Zhang – Ren – Sun 2016.
12 See Szegedy – Vanhoucke – Ioffe *et al.* 2016.
13 See Szegedy – Ioffe – Vanhoucke 2016.

Fig. 2: Extraction of rows of different lengths.

Classification on rows. The classification step was performed on the following arranged datasets:

- Dataset 1: 11 classes (Abraamios, Andreas, Dios, Dioscorus, Hermauos, Isak, Kyros 1, Kyros 3, Menas, Pilatos and Victor), 1232 pixels of row width;
- Dataset 2: 2 classes (Abraamios and Dios), 1232 pixels of row width;
- Dataset 3: 2 classes (Abraamios and Dioscorus), 1232 pixels of row width;
- Dataset 4: 2 classes (Abraamios and Dioscorus), 500 pixels of row width;
- Dataset 5: 4 classes (Abraamios, Dioscorus, Pilatos and Victor), 500 pixels of row width.

Dataset 1 obtained the worst performance because the classification task was more challenging, with an increasing number of classes to recognize. The best performance was obtained for dataset 2, particularly with the CNN InceptionResNetV2. This dataset comprised only two classes with a lot of samples for each. Neural Networks are known to work better with a huge and well-balanced dataset. Another interesting trend is that dataset 4 outperformed dataset 3 most of the time. Although they contain the same classes, dataset 4 contains 500 pixels-wide rows instead of 1232 pixels, so the number of images is higher for dataset 4. Dataset 5, with four writers, did not achieve a good performance.

Table 2: Accuracy achieved by classifying on rows.

Model	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
VGG19	35.98	86.65	63.91	80.03	42.75
ResNet50	38.17	97.68	67.68	79.85	44.19
InceptionV3	34.03	97.45	65.86	78.54	46.40
Inc.ResNetV2	42.07	98.35	66.81	48.07	42.55

Classification on Papyri Once the classification result was obtained for every row, it combining the predictions obtained from rows belonging to the same papyrus image was possible. The chosen combination rule was the Majority Vote, so if most of rows from a papyrus were classified as belonging to a certain class, then the entire papyrus

was classified as belonging to that class. The application of the combining rule involved only two of the datasets previously considered:

– Dataset 3: 2 classes (Abraamios and Dioscorus), 1232 pixels of row width;
– Dataset 4: 2 classes (Abraamios and Dioscorus), 500 pixels of row width.

Table 3: Accuracy achieved by classifying on rows.

Model	Dataset 3	Dataset 4
VGG19	67.86	46.43
ResNet50	76.79	83.93
InceptionV3	73.21	80.36
Inc.ResNetV2	65.48	88.10

The majority vote rule was applied on images belonging to the same papyrus to obtain a prediction on the entire papyrus instead of single fragments. This rule was applied to the third and fourth datasets, improving the classification result.

3.2 Second Approach

Image Pre-processing The image preprocessing step of the second approach starts from the greyscale images of the entire papyri, obtained during the first approach. We manually detected groups of two and then four consecutive characters from these. This procedure was repeated for every papyrus with the software LabelImg, which returned as output a .xml file containing the information regarding every ROI (Region Of Interest) detected on the image. A Python script received the images and the .xml files as input, returning all the patches containing two and four characters detected with LabelImg.

Classification on groups of two or four characters. As in the previous case, different datasets were generated to obtain balanced sets of two or four characters among the classes. A K-Fold Cross Validation strategy was adopted, and four CNN models, pre-trained on the public dataset ImageNet, were tested considering the DL techniques of Transfer Learning and Fine Tuning. The considered datasets are listed below:

– Dataset 1: 4 characters images, Abraamios and Dios;
– Dataset 2: 2 characters images, Dios vs. All;
– Dataset 3: 2 characters images, Dioscorus and Hermauos;
– Dataset 4: 2 characters images, Abraamios, Dios, Kyros and Victor.

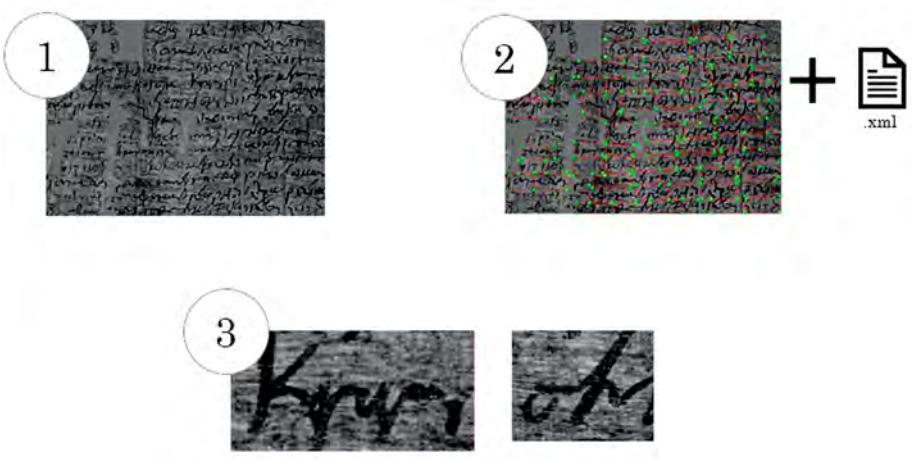


Fig. 3: Extraction of patches of two/four character from greyscale images, through LabelImg.

Table 4: Accuracy achieved by classifying on patches datasets.

Model	Dataset 1	Dataset 2	Dataset 3	Dataset 4
VGG19	59.12	60.36	51.72	52.3
ResNet50	94.58	95.08	73.71	58.8
InceptionV3	97.97	98.27	82.3	69.54
Inc.ResNetV2	98.59	97.69	85.48	63.75

The first dataset obtained the best performance, considering images containing four characters. The second dataset also achieved good results; we noticed that this was a common trend when one of the classes referred to Dios. The dataset that obtained the worst performance was the fourth.

Classification on Papyri. Once the classification result was obtained on every patch, combining the predictions obtained from patches belonging to the same papyrus image was possible. The chosen combination rule was the Majority Vote, applied to the following datasets:

- Dataset 1: majority vote on 4 characters images, Abraamios and Dios;
- Dataset 2: majority vote on 2 characters images, Dios vs. All;
- Dataset 3: majority vote on 2 characters images, Dioscorus and Hermauos;
- Dataset 4: majority vote on 2 characters images, Abraamios, Dios, Kyros and Victor.

Table 5: Accuracy achieved by classifying on patches datasets.

Model	Dataset 1	Dataset 2	Dataset 3	Dataset 4
VGG19	72.22	100	50	40.98
ResNet50	97.22	100	87.5	49.18
InceptionV3	94.44	100	75	59.01
Inc.ResNetV2	97.22	100	75	59.01

In the case of patches, the performance increased for the second dataset when applying a combination rule, but not for the others.

3.3 Third Approach

Image Preprocessing. The image preprocessing step of this approach was divided into two procedures: enhancement and binarization. Document enhancement involves improving the perceptual quality of document images and removing degradations and artifacts from the images. Document binarization separates each pixel belonging to the text from those belonging to the background. The enhancement is also a preprocessing step for binarizing degraded document images to remove unnecessary noise. The enhancing step was done according to the DeepOtsu method proposed by Sheng He and Lambert Schomaker, from the University of Groningen.[14] They built a network to improve input images by removing noise and correcting the degradation. Thus, the neural network's output was the improved version of the input with supervised learning. Many iterations could be performed according to the desired enhancing outcome. After several iterations, the output was the improved input version and could be binarized through classic binarization algorithms such as Otsu. The third approach required extracting patches of two characters from the original images, not the greyscale ones. The extraction process was the same as described for the second approach.

Classification This was a new approach, so the classification involved a dataset of only two writers: Hermauos and Isak. A K-Fold Cross Validation strategy was adopted, and four CNNs, pre-trained on the public dataset ImageNet, were tested considering the DL techniques Transfer Learning and Fine Tuning.

14 He – Schomaker 2019.

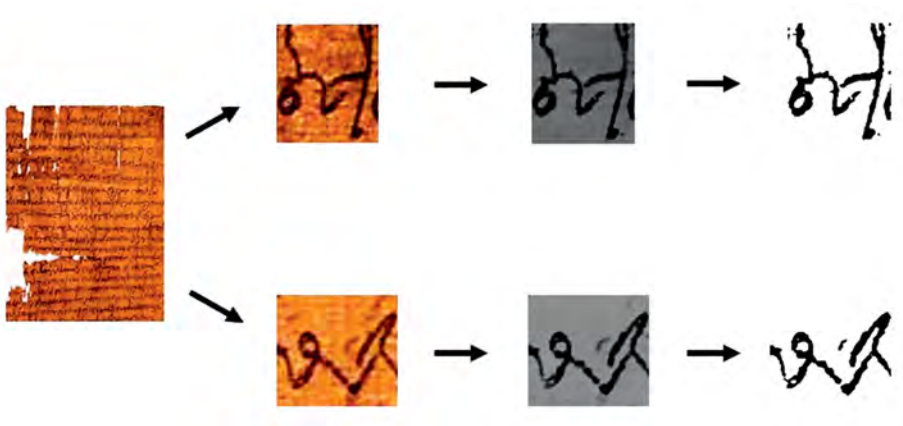


Fig. 4: Extraction of patches from original images, followed by enhancing and binarization.

Table 6: Accuracy achieved by classifying on binarized patches.

Model	Dataset
VGG19	74.42
ResNet50	74.53
InceptionV3	73.72
Inc.ResNetV2	72.56

The best classification result was obtained with the CNN ResNet50, though the results were almost identical for the other networks.

4 Conclusions

The study's primary objective was to develop a classification system based on deep learning to identify the authors of Greek papyri. This undertaking poses significant challenges due to various factors. One challenge is the limited quantity and quality of available data, which is essential for training AI models. Ancient papyri, the focus of this study, often suffer from constraints in data availability, as they do not cover a wide range of writers and stylistic variations. Moreover, preservation issues associated with these historical artifacts make it difficult to extract clear and accurate handwriting samples. Factors such as faded ink, smudges, tears, and other damage to the material can obscure the original script, complicating AI analysis.

However, despite these challenges, AI can be a valuable tool in authorship identification for papyri. Researchers can utilize methodologies involving image processing,

pattern recognition, and machine and deep learning algorithms to analyze existing data and make informed determinations. The initial findings from our study are promising and consistent with results from similar investigations. Future efforts will focus on several enhancements. Firstly, there will be a focus on integrating a balanced dataset, even if it requires removing numerous samples. A fine-tuning process will be implemented using a repository of handwritten texts. More binarization strategies will also be explored to eliminate background information and allow the network to concentrate solely on the personality of the writer's handwriting.

Bibliography

Christlein, V. – Marthot-Santaniello, I. – Mayr, M. – Nicolaou, A. – Seuret, M. (2022), *Writer Retrieval and Writer Identification in Greek Papyri*, in *Intertwining Graphonomics with Human Movements*, ed. by C. Carmona-Duarte – M. Diaz – M. A. Ferrer – A. Morales, Cham, 76–89, https://doi.org/10.1007/978-3-031-19745-1_6.

Cilia, N. D. – De Stefano, C. – Fontanella, F. – Marthot-Santaniello, I. – Scotto di Freca, A. (2021), *Papyrow: A Dataset of Row Images from Ancient Greek Papyri for Writers Identification*, in *Pattern Recognition. ICPR International Workshops and Challenges*, ed. by A. Del Bimbo – R. Cucchiara – S. Sclaroff – G. M. Farinella – T. Mei – M. Bertini – H. J. Escalante – R. Vezzani, Cham, 223–34, https://doi.org/10.1007/978-3-030-68787-8_16.

Cilia, N. D. – D'Alessandro, T. – De Stefano, C. – Fontanella, F. – Marthot-Santaniello, I. – Molinara, M. – Scotto Di Freca, A. (2024), *A Novel Writer Identification Approach for Greek Papyri Images*, in *Image Analysis and Processing – ICIAP 2023 Workshops (Udine, September 11–15, 2023) Proceedings*, ed. by A. Del Bimbo – R. Cucchiara – S. Sclaroff – G. M. Farinella – T. Mei – M. Bertini – H. J. Escalante – R. Vezzani, Heidelberg, II, 422–36, https://doi.org/10.1007/978-3-031-51026-7_36.

Deng, J. – Dong, W. – Socher, R. – Li, L.-J. – Li, K. – Fei-Fei, L. (2009), *Imagenet: A Largescale Hierarchical Image Database*, in *2009 IEEE Conference on Computer Vision and Pattern Recognition (Miami, 20–25 June 2009)*, Piscataway, 248–55, https://doi.org/10.1109/CVPR.2009.5206848.

Fontanella, F. – Colace, F. – Molinara, M. – Scotto Di Freca, A. – Stanco, F. (2020), *Pattern Recognition and Artificial Intelligence Techniques for Cultural Heritage*, Pattern Recognition Letters 138, 23–9, https://doi.org/ 10.1016/j.patrec.2020.06.018.

Fournet, J.-L. (2008), ed., *Les archives de Dioscore d'Aphrodité cent ans après leur découverte. Histoire et culture dans l'Egypte byzantine*, Paris.

Gupta, A. – Kumar, S. – Gupta, R. – Chaudhury, S. – Joshi, S. (2007), *Enhancement of Old Manuscript Images*, in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Piscataway, II, 744–8, https://doi.org/10.1109/ICDAR.2007.4377014.

He, S. – Schomaker, L. (2019), *DeepOtsu: Document Enhancement and Binarization using Iterative Deep Learning*, Pattern Recognition 91, 379–90, https://doi.org/10.1016/j.patcog.2019.01.025.

He, K. – Zhang, X. – Ren, S. – Sun, J. (2016), *Deep Residual Learning for Image Recognition*, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (Las Vegas, 27–30 June 2016)*, Piscataway, 770–8, https://doi.org/10.1109/CVPR.2016.90.

Jayanthi, N. – Indu, S. – Hasija, S. – Tripathi, P.: *Digitization of Ancient Manuscripts and Inscriptions: A Review*, in *Advances in Computing and Data Sciences: First International Conference, ICACDS 2016, Ghaziabad, India, November 11–12, 2016, Revised Selected Papers*, Cham, I, 605–12, https://doi.org/10.1007/978-981-10-5427-3_62.

Mohammed, H. – Marthot-Santaniello, I. – Märgner, V. (2019), *GRK-Papyri: A Dataset of Greek Handwriting on Papyri for the Task of Writer Identification*, in *2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia*, Piscataway, 726–31, https://doi.org/10.1109/ICDAR.2019.00121.

Nasir, S. – Siddiqi, I. (2021), *Learning Features for Writer Identification from Handwriting on Papyri*, in *Pattern Recognition and Artificial Intelligence*, ed. by C. Djeddi – Y. Kessentini – I. Siddiqi – M. Jmaiel, Cham, 229–41, https://doi.org/10.1007/978-3-030-71804-6_17.

Nasir, S. – Siddiqi, I. – Moetesum, M. (2021), *Writer Characterization from Handwriting on Papyri Using Multi-Step Feature Learning*, in *Document Analysis and Recognition – ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings*, Berlin – Heidelberg, I, 451–65, https://doi.org/10.1007/978-3-030-86198-8_32.

Peer, M. – Sablatnig, R. (2023), *Feature Mixing for Writer Retrieval and Identification on Papyri Fragments*, in *7th International Workshop on Historical Document Imaging and Processing, San Jose, CA, USA, August 2023*, New York, https://doi.org/10.1145/3604951.3605515.

Rehman, A. – Naz, S. – Razzak, M. I. (2019),: *Writer Identification Using Machine Learning Approaches: A Comprehensive Review*, Multimedia Tools and Applications 78, 10889–931, https://doi.org/10.1007/s11042-018-6577-1.

Simonyan, K. – Zisserman, A. (2015), *Very Deep Convolutional Networks for Large-Scale Image Recognition*, https://arxiv.org/abs/1409.1556.

Szegedy, C. – Ioffe, S. – Vanhoucke, V. (2017), *Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning*, in *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (San Francisco, February 4–9, 2017)*, ed. by S. Singh – S. Markovitch, https://dl.acm.org/doi/10.5555/3298023.3298188.

Szegedy, C. – Vanhoucke, V. – Ioffe, S. – Shlens, J. – Wojna, Z. (2016), *Rethinking the Inception Architecture for Computer Vision*, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (Las Vegas, 27–30 June 2016)*, Piscataway, 2818–26, https://arxiv.org/abs/1512.00567.

Indices

All along the volume, the standard bibliographical references to papyrus editions (after the online Checklist of Editions at http://papyri.info/docs/checklist) are used. For the abbreviations of digital resources see the appropriate index. Bibliographical abbreviations of academic journals follow the Année Philologique.

I Digital resources and projects

https://doi.org/10.1515/9783111070162-021

II Modern scholars

III Ancient people

IV Papyrus texts

V Literary sources