Virtual Research Environments

Caroline T. Schroeder

https://orcid.org/0000-0001-9543-0692

Abstract Virtual Research Environments in theological studies (and esp. early Christian studies and the related field of Classical studies) can provide valuable infrastructure for producing digital editions of primary sources and enabling other forms of digital and computational research. Creating and sustaining these environments has challenges. This chapter examines the benefits of collaborating across projects as well as sharing and reusing digital resources. The chapter also presents some of the considerations for working with *messy* or *clean* digital data, and for adopting existing technical standards. With respect to all of these issues, building and using VREs involves developing relevant technical infrastructure. But just as important as technology are the humanistic questions and collaborative personal relationships underpinning a successful digital initiative.

Keywords Digital Humanities, Virtual Research Environments, Tools, Standards, Collaboration, Open Access, Data Cleaning, Early Christian Studies

1. Introduction

Virtual Research Environments (VRE) in theological studies (and esp. early Christian studies and the related field of Classical studies) can provide valuable infrastructure for producing digital editions of primary sources and enabling other forms of digital and computational research. Creating and sustaining these environments has challenges. Some key elements of successful VREs include collaboration across different projects, sharing and reusing digital resources, and careful consideration of how to work with *messy* or *clean* digital data, and whether to adopt existing technical standards. In this paper, I will address these aspects of Digital Humanities work in our field through the lens of the creation of the *Coptic Scriptorium* (CS) platform. While the focus of this essay is CS, other VREs are considered and the analysis extends beyond the scope of our individual experience.

The virtual research environment I co-direct, *Coptic Scriptorium*, originated at a *National Endowment for the Humanities* summer research institute hosted by the *Perseus Digital Library* at Tufts University in 2012. Researchers from all career stages – from graduate student to full professor – working in a variety of languages – Greek, Latin, Russian, Coptic – applied and attended a three-week workshop co-directed by

Caroline T. Schroeder: Virtual Research Environments, in: Christopher A. Nunn and Frederike van Oorschot (eds.): Compendium of Computational Theology 1. Introducing Digital Humanities to Theology, Heidelberg: heiBOOKS, 2024, S. 379–392. DOI: https://doi.org/10.11588/heibooks.1521.c21960

Monica Berti (a Classicist and Digital Humanist, now of Leipzig University), Gregory Crane (Tufts University, *Perseus* founder), and Anke Lüdeling (Corpus Lingustics, Humboldt University). At that point, "digital Coptic" was in its infancy, and there were few openly accessible VREs for early Christian studies or ancient studies. The *Perseus Digital Library*[1], our institutional host, was one of the most well-known (Crane 1998). *Trismegistos*[2] served as a *linked-data* hub for people, places, and ancient texts (building on and collaborating with the *Heidelberger Gesamtverzeichnis griechischer Urkunden aus Ägypten* [HGV] and the *Leuven Database of Ancient Books* [LDAB]) (Depauw & Gheldof 2014). *Papyri.info* created a cutting-edge collaborative text-editing environment that benefitted from crowd-sourcing among papyrologists.[3] The *Tesserae Project* at the University of Buffalo had also launched, facilitating research in text-reuse in classical sources (Forstall et al. 2011; Okuda et al. 2022; cf. also the chapter by J. Nantke in this volume, p. 288). While there were additional subscription-based research environments for Greek and Latin, open access or open-source environments were few – those aforementioned being some of the main projects. The organizers of the NEH Institute hoped its participants would be inspired to fill the gaps.

In Coptic Studies, the Unicode character set for the Coptic alphabet[4] had been approved in 2004, with major additions in subsequent years, including important diacritics such as the binding macron characters and the "binding ni" that appears at the end of lines in manuscripts in 2007.[5] *Papyri.info* had recently begun publishing some Coptic papyri and ostraca. Other institutes and individuals had been working on both Coptic and Syriac texts non-Unicode fonts and circulating digital forms of the New Testament and Christian Old Testament in these languages (Schroeder 2019). Additionally, Tito Orlandi's work at the *Corpus dei Manuscritti Copti Letterari* (CMCL), ongoing for decades, was foundational (Orlandi 1997a; b; 2021). Nonetheless, sustainable digital editions of Coptic literature and sustained digital and computational research in Coptic studies were only at the beginning stages. Amir Zeldes, a linguist at Humboldt University (not a "Coptologist") and I (not a linguist), met at the Tufts NEH Institute and discovered our shared interest in Coptic literature and Digital Humanities, and began planning the project. *Coptic Scriptorium*[6] launched its first

[1] See http://www.perseus.tufts.edu (Accessed: 25 June 2024).

[2] See http://www.trismegistos.org (Accessed: 25 June 2024).

[3] See http://papyri.info/ddbdp (Accessed: 25 June 2024).

[4] See https://www.unicode.org/wg2/docs/n2824.pdf (Accessed: 25 June 2024).

[5] For the revisions in 2004, see the worksheet under https://www.unicode.org/wg2/docs/n2744.pdf; for the revisions in 2007 see http://unicode.org/wg2/docs/n3222 and https://www.unicode.org/L2/L2007/07118.htm [Protocol of the UTC 111/L 2 208 Joint Meeting]. For the standard Unicode font *Antinoou* (2012) see http://www.evertype.com/fonts/coptic. All addresses were accessed on 25 June 2024.

[6] See https://copticscriptorium.org (Accessed: 25 June 2024).

pilot corpus, natural language processing tools, and one-page website in 2013.[7] We now have a database of Coptic literature of over 1.2 million words (annotated for part of speech, syntax, entities, lemmas, language of origin, manuscript information, and more), plus multiple tools including an online natural language processing pipeline (Schroeder & Zeldes 2013–2023; 2016; 2020).

In this paper, I will address three key issues in developing VREs that have posed both challenges and opportunities as our project has grown over the past decade: specialization and collaboration in reuse of data and tools, messy data, and technical standards. While building and using VREs for Digital Humanities research involves developing technical infrastructure, just as important are pursuing the humanistic questions and collaborative personal relationships underpinning a successful digital initiative.

2. Specialization, Collaboration, and Reuse

Digital research environments are expensive endeavors, and often the audience or user-community for such environments is small. In Coptic Studies, for example, most of us know each other, whether we work in North America, Europe, Australia, Egypt, or Japan. And there is little room for overlap in research – if we already know someone is working on an edition of certain manuscripts or papyri, the rest of us usually go off to work on something else. This also has proven particularly true in digital Coptic Studies. The ecosystem that has emerged consists of specialists in specific areas. And while early Christian studies, Classics, and Biblical Studies have more expansive communities, in the digital realm, the costs of creating VREs discourages duplication. Thus, in textual and linguistic Coptic studies, the major open access projects specialize in different aspects of the field. Each of these research environments has developed in response to the particular research needs of a certain research community, and each has limitations as well as benefits.

Papyri.info publishes digital editions of ostraca and papyri using the XML *(extensible markup language)* standards created by the *Text Encoding Initiative* and the EpiDoc subcommunity of the TEI (Elliott et al. 2006–2021).[8] Papyri and ostraca tend to be shorter than literary texts, and *Papyri.info* creates a digital research environment comparable to the analog research methods papyrologists traditionally have employed (editions and translations with notes, images, apparatus, etc.). As a result,

[7] Although we no longer have a copy of the original website, the version from 9 October 2014 is archived at the Internet Archive Wayback Machine: https://web.archive.org/web/20141009102742/http://www.copticscriptorium.org (Accessed: 25 June 2024).

[8] See on TEI: http://www.tei-c.org; see on EpiDoc: http://epidoc.stoa.org. Both addresses were accessed on 25 June 2024.

crowd-sourcing digitization of papyri among papyrologists has been feasible. Certainly, *Papyri.info* has invested a significant amount of time and resources in training and outreach, which cannot be understated; the genre of sources and the digital methods also contribute to its success in publishing a tremendous number of documents. There are some features, however, that this environment either does not have or has some challenges with (None of the following observations should be taken as criticisms – it is a remarkable achievement in scope and method. The description of the platform's parameters exemplifies how this particular VRE serves specific research questions and methodologies.). While the platform does enable searching of individual words and series of words (including using regular expressions) and provides rich, searchable metadata, users I have met at conferences sometimes express concern that the results may omit some hits or that they are not sure how to use the interface to produce searches as comprehensive as they would like. Downloading results for computational work is challenging for a basic user, and the words are not linked to an online dictionary as in the *Perseus Digital Library. Papyri.info* is a crown jewel of digital ancient studies, because it provides the features that it *does* very well. No one platform can do all things for all users.

Similarly, we see specialization (and thus different features) in other open-access VREs. The *Göttingen Old Testament Project*[9] produces digital editions of Coptic Christian Old Testament manuscripts utilizing the *Virtual Manuscript Room* environment created originally by the *Institut für Neutestamentliche Textforschung*, also using TEI markup (Behlmer 2017). The PAThs project in Rome has created an *archaeological atlas of Coptic literature*[10] by building an information hub about literary manuscript data – where codices were produced and found, where they are now archived or stored, where they have been published, which works are preserved on each codex, etc. (Buzi 2017; Buzi et al. 2018). The *Thesaurus Linguae Aegyptiae* (in collaboration with others) published an Egyptian Coptic lexicon formatted in TEI-XML, which CS then instrumentalized into an online dictionary[11], and the *Database and Dictionary of Greek Loanwords* in Coptic project subsequently contributed their Greek lemma list and definitions (Feder et al. 2018; Burns et al. 2019).

Collaborating with other projects or reusing their open-source data or technology also enables projects to excel in their own areas of research without having to reinvent the wheel in others. For the most part, digital papyrological projects collaborate with *Papyri.info* so that their data feeds into the common shared database. This allows for institutions with papyri collections to focus on their specific items while also contributing to a shared resource benefitting a wider scholarly community.

[9] See https://www.uni-goettingen.de/en/digital+edition+des+koptischen+(sahidischen)+alten+Testaments/475974.html (Accessed: 25 June 2024).

[10] See https://atlas.paths-erc.eu (Accessed: 25 June 2024).

[11] See https://coptic-dictionary.org/about.cgi (Accessed: 25 June 2024).

The *Coptic Dictionary Online* (CDO) presents another example of reuse and collaboration among specialists. It contains the lexica from two projects, being the *Dictionary and Database of Greek Loan Words in Coptic* and the *Thesaurus Linguae Aegyptiae*. The CDO links each dictionary entry to individual words in the corpora published in CS's database; similarly, the CS database links back to the CDO word by word. In addition, entries for Egyptian Coptic words link to an online pdf of the most comprehensive print Coptic dictionary (by Crum [1939], hosted by yet another partner, the *Göttingen Old Testament project*). Entries for Greek loanwords link to the *Perseus* online Greek dictionary. The CS-team created and maintains the online interface that enables searching of the CDO and all of the interlinking of resources. Such a comprehensive, interlinked resource used widely internationally could not have been accomplished by one research unit alone.

Such accomplishments do not come without challenges, however. In Coptic, for example, Coptologists differ on what constitutes a word in the language. This may sound arcane, but the issue directly affects the creation of an online dictionary. Coptic is an agglutinative language, which means that different linguistic units (such as a subject pronoun and a verb) are bound together and are written together; additionally Coptic manuscripts are written in *scriptua continua*, with no white space between words or bound groups of words. Word segmentation is important for search, and also for creating lexical resources, such as a dictionary. Take the term for "idol-worshipper," *refšmšeeidolon*. Should we treat this term as one word with one lexical entry, since linguistically the whole item is a noun that takes one definite or indefinite article and as one term can be a subject of a verb? Or should we treat it as three words based on the morphemes that combine to create the term *(ref-šmše-eidolon)*? Where *šmše* means "to worship," *eidolon* is "idol", and *ref* is the prefix that indicates a term is a noun in the form of "the person who" does the thing that follows (the person who worships idols, or "idol-worshipper"). CS – with our interest in linguistics, part of speech annotation, and syntax annotation – treats the term as one word (a noun) with three morphemes. The TLA's research interests in creating their Egyptian Coptic lexicon concern (in part) tracing the Egyptian language through all its phases. Thus, it treats *ref-* as its own lexical entity as a lemma and gives it an entry in the *Coptic Dictionary Online* ("TLA lemma no. C3102“). Clicking on the link within that entry to find instances of the "word" *ref-* in the CS database, however, does *not* result in hits for all instances of *ref-* in our corpora, since we treat this morpheme as a prefix and not a lemma or word in and of itself; the query linking the CDO and CS corpora database is automated, so the different data models lead to a bit of a mismatch in a small number of cases (such as the morpheme *ref-*).

Deciding on a common definition of what constitutes a Coptic word or lemma before launching the CDO would have ground this collaboration to a halt. Instead, the projects agreed that some inconsistencies in mapping across our data were a small price to pay for the overall benefit of linking the dictionary to an online database of Coptic textual corpora. Sometimes these inconsistencies can be resolved in at least one direction; in the CS database, we do annotate a word like *refšmšeeidolon* as three morphemes, with a link for each morpheme to that morpheme's entry in the online dictionary. One might not be able to get to all the CS database hits for words that begin with *ref-* with one click from the CDO entry, but one can get them with a slight manual modification to the database query language. Additionally, one can access the dictionary entry for *ref-* with one click from the CS database. Manual mapping of entries alleviates some other inconsistencies, but such coding requires human labor, which can be challenging given the competitive and meagre funding opportunities for many humanities projects.

The very beginning of CS also benefitted from significant use of prior work, including open-source technology. Tito Orlandi's lexicon (published at CMCL) enabled us to create natural language processing tools that segmented Coptic text into words and tagged them with their parts of speech within the first year of the project. It easily cut a year off our initial work time. Instead of building our own database infrastructure, we adapted an open-source tool developed by linguists (including CS co-founder Zeldes) (Zeldes et al. 2009; Krause & Zeldes 2014). Again, this reuse allowed us to publish a searchable corpus of texts within months, not years. On the other hand, philologists and historians who are unfamiliar with corpus linguistics as a method can find the tool's search interface challenging. As a result, we have posted online tutorials and cheat-sheets to help users navigate the system and have invested development resources into modifying the tool for Coptic. While not perfect, the benefits of a robust, nearly out-of-the-box infrastructure outweigh the drawbacks, and certainly outweigh the costs of building an entirely new database infrastructure from scratch.

By necessity, I have not included all VREs for ancient studies or early Christian studies in this discussion of collaboration and reuse. Nonetheless, these examples illustrate some of the challenges resulting from specialization, disciplinary diversity, and intra-disciplinary methodological differences. Moreover, despite these challenges, open-source and open-access VREs administered by projects open to collaboration and data-sharing can stimulate much more robust research opportunities than more siloed endeavors.

3. Messy vs Clean Data

One interdisciplinary debate within Digital Humanities that directly affects VREs in ancient studies and early Christian studies is the degree to which we should clean our textual data. Philology as a discipline prizes accuracy and precision in text editions as well as in translations. Corpus linguists, computational linguists, and some digital humanists have a higher tolerance for mess.

"Messy" humanities data traditionally has been understood as large quantities of unstructured and unedited text (*big data*, Schöch 2013). Until recent years, scholars working in antiquity have not even had access to “big” ancient textual data in digital form. For Greek and Latin, *Perseus,* and *Open Philology* especially, but other projects as well, have contributed to large-scale digitization. For Coptic, Syriac, Ge’ez, and other languages, we are slowly moving towards what we might call *medium data.* Digital ancient studies experiences a push-and-pull between the desire for larger corpora of digital data we can search or analyze on the one hand, and the prioritizing of highly accurate, thoroughly peer reviewed editions on the other. In a 2013 conference paper about the creation and long-term viability of *Papyri.info,* Roger S. Bagnall cited the peer-review process as one of the factors slowing down the process of publishing more digital editions in their platform. Much of *Papyri.info* replicated in the digital realm – albeit in transformed ways – the scholarly form papyrologists were used to producing and using – the edition. And as such, it developed a peer review process before publishing editions online, much as print editions undergo peer review. The backlog of papyri or ostraca awaiting publication accumulated to the point at which the project board decided to publish editions that had not yet undergone the final round of peer review (Bagnall 2013). In digital publications, of course, we can release a new version with any corrections or editorial emendations quickly. In traditional print publications of editions and translations, scholars may labor for a decade or more ensuring the text is accurate with detailed apparatus notes, or commentary; except for extremely commonly read works, the appearance of revised editions or new editions by other scholars soon after the previous publication is rare. *Perseus* founder Gregory Crane commented on this phenomenon back in the 1980s in an early paper on Classics and “hypertext” (Crane 1987).

In a digital age, *messy* data might mean a variety of things. Inaccuracies in optical character recognition (OCR) when digitizing print editions. Typographical errors in transcriptions of ancient texts. Typographical errors in metadata. Misattribution of sources or inaccuracies in dating. For text with annotations for linguistic information such as part of speech, links to other resources, manuscript information, etc., any annotation errors also render data *messy.* Scholars editing, translating, and interpreting ancient texts often express that we are accustomed to working with highly accurate editions on all of these measures – accuracy of text, information about the work containing the text, translation, and so forth. The reality is that we find errors in print editions, as well. Our tolerance for mess, however, may be lower than what is required for working with automated digital methods. Accuracy rates of 98–99% for OCR, e.g., are considered quite high; in a million-word corpus, such a rate means ten to twenty thousand characters are affected – a number to which corpus linguists or computer scientists might be accustomed but that many philologists may find troubling (on OCR for historical languages generally, see Smith & Cordell 2018).

Some digital humanists have recently published work advocating for more tolerance for mess. Mess can involve *inaccuracies* in data or challenges to highly structured, formal systems and ideologies underpinning some computational work. In the latter understanding, as Losh et al (2016) have written, “‘Mess’ serves as a theoretical intervention in popular notions of digital media as neat, clean, and hyper-rational." Similarly, Katie Rawson and Trevor Muñoz argue that the debate over clean vs. messy data is an epistemological one: "The term 'cleaning' implies that a dataset begins as 'messy.' 'Messy' suggests an underlying order: it supposes things already have a rightful place, but they are not in it – like socks on the bedroom floor rather than in the bureau or the hamper" (Rawson & Muñoz 2019). In this view, cleaning a dataset – esp. normalizing, or annotating to create a structured data*set* out of unstructured *data* – involves imposing a pretheorized or presupposed order or model on the data. "The cleaning paradigm assumes an underlying, 'correct' order." Rawson & Muñoz (2019) advocate for embracing the diversity of unstructured data, and for allowing the querying and discovery of *uncleaned* data to point us to new understandings of the data and the communities that gave birth to it.

In philology – and here I specifically refer to ancient literature, especially biblical studies, rather than papyrology – the quest for *clean* textual data is connected to the quest for the *urtext. Clean,* here, is not perfectly spelled or accurately annotated text, but the earliest version of the work, the one closest to the original. Often the cleanest critical edition of a work matches no known manuscript 100 %. VREs and methodologies in manuscript studies take two different approaches to this pre-digital methodology. Tools and projects sometimes digitally replicate this traditional process, by transcribing (or creating VREs for transcribing) manuscript witnesses that will then be compared digitally to produce a critical edition. (Behlmer 2017; Huskey 2019) Tools such as *Juxta Commons* and *CollateX*[12] allow researchers to mark up parallel witnesses of the same text during the digital editing process (Wheeler & Jensen 2014).

Some digital humanists in Classics have also investigated how to produce digitally the print apparatus philologists are used to seeing; as a "data visualization," the apparatus is efficient and effective (Fischer 2019; Huskey 2022). Other projects, such as CS, publish digital editions of manuscript transcriptions (as well as earlier print editions) with metadata connecting versions of the same work to each other but without producing an apparatus or critical edition. In this way, at least, CS has embraced "mess." We certainly impose order on the text through our linguistic annotations, which employ a data model based in large part on the grammatical categories and syntax in Bentley Layton's *Coptic Grammar* – a work itself critiqued for aggressively creating and imposing new linguistic categories (Layton 2011; Shisha-Halevy 2006; Feder 2017). But with respect to the editions of Coptic literature, when publishing transcriptions of manuscripts, we transcribe the original text (however *messy*) and produce normalized and lemmatized text (the more *cleaned* textual data) as annotations on the original. Thus, the researcher may search for an expected, *cleanly* spelled word and also see all the instances of that term in its original spelling in our database. One can also pull up parallel manuscript witnesses, in the cases where we have published them. But we provide no critical edition or apparatus.

[12] See https://collatex.net/about (Accessed: 25 June 2024).

4. Technical Standards

Traditionally in Digital Humanities, technical standards have provided three important functions. Standards lay the groundwork for how to mark up or process data so that subsequent projects do not have to reinvent the wheel. In this way, they provide a shared resource for humanists working in related research areas. To my mind, this is the most important aspect of digital standards – a community comes together to create a roadmap for each other and for researchers of the future. Even if not all aspects of set of standards fit every individual project in a field, the standards provide a starting point. Additionally, they will point other researchers toward known issues in digitization or computation in their scholarly field. For example, the PAThs project's data model includes more than one field for the author of a work – the "stated" author (as stated in the manuscript or work) and the "creator" (the verifiable historical author) (Buzi et al. 2018). Studying their data model and standards can help any project working on manuscripts and historical literature.

Standards also in theory help ensure consistency of data and annotations. For example, geographical locations annotated in the same way across a dataset allow researchers to query for a place and have a reasonable expectation of finding most if not all the instances of that location. Different textual data annotated according to the same standard by multiple projects also can be queried and analyzed in a comparative way. One such example is the *Universal Dependency dataset* (UD), in which corpora from over 100 languages have been annotated according to the same linguistic standards. Although Coptic has long been considered an *under-resourced* and perhaps even obscure language, its presence in the UD means that researchers have examined it alongside modern tongues, such as Danish and Chinese, to gain insight into language (Zeldes & Abrams 2018; Pinter et al. 2019; Chen et al. 2022).

Finally, this consistency in theory should lead to more interoperability between projects and research environments. Digital editions marked up according to a shared standard (such as TEI-XML) in one VRE should be publishable or editable in another VRE using the same standards. *Papyri.info* provides such an example; it aggregates papyri and ostraca digitized by multiple projects in one platform, a process possible in part due to shared use of the EpiDoc subset of the TEI-XML standards.

In practice, however, annotation is a process of interpretation. How to implement the same standard can vary. CS, the *Göttingen Old Testament* Project, and the *Canons of Apa John* project all have agreed to data share. We all use TEI-XML to annotate for manuscript information in our diplomatic transcriptions. However, the use some of the XML tags in slightly different ways, and we also have different understandings of what binds Coptic words into phrases called "bound groups." As a result, we created converter scripts to ensure true interoperability. These differences do not impose critical or unresolvable obstacles to collaborations, but they do point to the human element in data sharing. Additionally, interdisciplinary projects may find that no one set of standards can capture all the information their project will digitize and annotate. For example, CS releases our data in multiple formats according to different standards, because these standards have developed within specific fields for their individual disciplinary needs and research questions. While TEI-XML provides a robust tagset for digital editions, annotating for part of speech and syntax requires different kinds of annotations. Thus, our project releases our annotated corpora in multiple formats; each document is released as a *light* single TEI-XML file that captures manuscript information and some basic linguistic information (language of origin, lemma, part of speech), PAULA XML documents with full stand-off annotations for all aspects of our data model (including codicological and linguistic annotations), relational database files that contain full metadata and textual annotations used to populate the ANNIS database for querying of our corpora, and a SGML document with all annotations and metadata contained in one file.[13] We generate the files in these different formats from one master file. Moreover, we release the aforementioned UD corpus, which is a subset of our corpora with a high level of accuracy annotated according to the UD treebank syntax standards.

Communication and commitment to collaboration within disciplines and across disciplinary differences are just as important as technical standards. Such communication also extends beyond the scope of documentation. Documentation has long been raised as a key feature for sustainability and useability in Digital Humanities projects. It is also a common challenge, especially for projects running on limited funding and/or an abbreviated timeline for funding (Edmond & Morselli 2020). A project's standards as well as the decision-making process or technical investigations behind those standards can – and should – be documented in journal articles, project blogs, white papers, and "Read Me" files. Transparency about how a VRE functions, why it functions that way, and who contributed to the labor of the project is important (Keralis et al. 2023). In small research fields, successful projects cultivate a human mindset of collaboration and ongoing communication with users and research partners alongside providing documentation of standards.

5. Conclusion

Many discussions of VREs or other *tools* in Digital Humanities center on questions about sustainability (cf. the chapter by J. Apel in this volume, pp. 402–403). In building a tool or platform, project teams must consider the human labor required for creating it and supporting it over time, esp. as technology and standards change. VRE-teams need to consider how to provide sufficient training and documentation for users. Sustainability is also a human question as much as a technical one. Developing a VRE that is flexible enough to survive beyond initial startup funding (or to produce

[13] See https://github.com/CopticScriptorium/corpora (Accessed: 25 June 2024).

data in formats that survive) requires both technical expertise and personal commitments to such an approach. The topics I have addressed in this essay are embedded in conversations about Digital Humanities sustainability. Projects using VREs in Digital Humanities can benefit from considering how they might reuse existing data and tools – thus extending the lifecycle of other projects' output and possibly reducing the financial cost of the labor of development in their own projects. Conversations about technical standards and *messy* or *clean* data are essential when developing plans for sunsetting a project. Planning for collaboration at the outset can help projects avoid "reinventing the wheel" and also can enable use of their data or tools on a wider scale and longer timeline. Although a VRE is technical infrastructure, the questions, and methods essential to building and maintaining such a tool are deeply human.

References

Bagnall, R.S. (2013). Digital Presentation, Digital Editing, Digital Community. The Case of Papyrology. In *Meeting Abstracts. SBL Meeting 2013*. Baltimore: Society of Biblical Literature. URL: https://www.sbl-site.org/meetings/Congresses_Abstracts.aspx?MeetingId=23 (Accessed: 25 June 2024).

Behlmer, H. (2017). Die digitale Gesamtausgabe und Übersetzung des koptisch-sahidischen Alten Testaments. Ein neues Forschungsprojekt an der Akademie der Wissenschaften zu Göttingen, *Early Christianity,* 8(1), 97–107. DOI: https://doi.org/10.1628/186870317X14876711440169 (Accessed: 25 June 2024).

Burns, D.M., Feder, F., John, K., & Kupreyev, M. (2019). *Comprehensive Coptic Lexicon. Including Loanwords from Ancient Greek* [data set]. DOI: https://doi.org/10.17169/REFUBIUM-2333 (Accessed: 25 June 2024).

Buzi, P. (2017). Tracking Papyrus and Parchment Paths. An Archaeological Atlas of Coptic Literature. Literary Texts in Their Geographical Context. Production, Copying, Usage, Dissemination and Storage (PAThs), *Early Christianity,* 8(4), 507–516. DOI: https://doi.org/10.1628/186870317X15100584934630 (Accessed: 25 June 2024).

Ead., Bogdani, J., & Berno, F. (2018). The 'PAThs'-Projekt. An Effort to Represent the Physical Dimension of Coptic Literary Production (Third-Eleventh Centuries), *Comparative Oriental Manuscript Studies Bulletin,* 4(1), 39–58. DOI: https://doi.org/10.25592/uhhfdm.253 (Accessed: 25 June 2024).

Chen, X., Gerdes, K., Kahane, S., & Courtin, M. (2022). The Co-Effect of Menzerath-Altmann Law and Heavy Constituent Shift in Natural Languages. In M. Yamazaki, H. Sanada, R. Köhler, Sh. Embleton, R. Vulanović & E.S. Wheeler (Eds.), *The Co-Effect of Menzerath-Altmann Law and Heavy Constituent Shift in Natural Languages* (pp. 11–24). Mouton: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110763560-002 (Accessed: 25 June 2024).

Crane, G. (1987). From the Old to the New. Intergrating Hypertext into Traditional Scholarship. In *Proceedings of the ACM Conference on Hypertext* (pp. 51–55). New York: Association on Computing Machinery. DOI: https://doi.org/10.1145/317426.317432 (Accessed: 25 June 2024).

Id. (1998). The Perseus Project and Beyond. How Building a Digital Library Challenges the Humanities and Technology, *D-Lib Magazine,* no pag. URL: http://www.dlib.org/dlib/january98/01crane.html (Accessed: 25 June 2024).

Crum, W. E (1939). *Ein koptisches Wörterbuch.* Oxford: Clarendon Press.

Depauw, M., & Gheldof, T. (2013). Trismegistos. An Interdisciplinary Platform for Ancient World Texts and Related Information. In Ł. Bolikowski, V. Casarosa, P. Goodale, N. Houssos, P. Manghi, & J. Schirrwagen (Eds.), *Theory and Practice of Digital Libraries. TPDL 2013. Selected Workshops.* Cham: Springer [= *Communications in Computer and Information Science,* 416]. DOI: https://doi.org/10.1007/978-3-319-08425-1_5 (Accessed: 25 June 2024).

Edmond, J., & Morselli, F. (2020). Sustainability of Digital Humanities Projects as a Publication and Documentation Challenge, *Zeitschrift für Dokumentation,* 76, 1019–1031.

Feder, F. (2017). Review of *A Coptic Grammar* by Bentley Layton, *Orientalistische Literaturzeitung,* 112(2), 108–12. DOI: https://doi.org/10.1515/olzg-2017-0035 (Accessed: 25 June 2024).

Id., Kupreyev, M., Manning, E., Schroeder, C. T., & Zeldes, A. (2018). A Linked Coptic Dictionary Online. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 12–21). Santa Fe, New Mexico: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W18-4502 (Accessed: 25 June 2024).

Fischer, F. (2019). Digital Classical Philology and the Critical Apparatus. In M. Berti (Ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution* (pp. 203–220). Berlin/Boston: De Gruyter Saur. DOI: https://doi.org/10.1515/9783110599572-012 (Accessed: 25 June 2024).

Forstall, Ch. W., Jacobson, S. L., & Scheirer, W. J. (2011). Evidence of Intertextuality. Investigating Paul the Deacon's *Angustae Vitae, Literary and Linguistic Computing,* 26(3), 285–296. DOI: https://doi.org/10.1093/llc/fqr029 (Accessed: 25 June 2024).

Huskey, S. (2019). The Digital Latin Library. Cataloging and Publishing Critical Editions of Latin Texts. In M. Berti (Ed.), *Digital Classical Philology* (pp. 19–34). Berlin/Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110599572-003 (Accessed: 25 June 2024).

Id. (2022). The Visual [Re]Presentation of Textual Data in Traditional and Digital Critical Editions, *Magazén,* 1. DOI: https://doi.org/10.30687/mag/2724-3923/2022/05/005 (Accessed: 25 June 2024).

Keralis, S. D. C., Mirza, R., & Seale, M. (2023). Librarians' Illegible Labor. Toward a Documentary Practice of Digital Humanities. In M. K. Gold & L. F. Klein (Eds.), *Debates in the Digital Humanities* 2023 (no pag.). Minneapolis: University of Minnesota Press. URL: https://dhdebates.gc.cuny.edu/read/debates-in-the-digital-humanities-2023/section/c8bfbcfa-1500-41c2-a1d7-63b8c81b627f#ch20 (Accessed: 25 June 2024).

Krause, Th., & Zeldes, A. (2014). ANNIS3. A New Architecture for Generic Corpus Query and Visualization, *Digital Scholarship in the Humanities,* 31(1), 118–139. DOI: https://doi.org/10.1093/llc/fqu057 (Accessed: 25 June 2024).

Layton, B. (2011). *A Coptic Grammar.* 3. ed. Wiesbaden: Harrassowitz [= *Porta Linguarum Orientalium. Neue Serie,* 20].

Losh, E., Wernimont, J., Wexler, L., & Wu, H.-A. (2016). Putting the Human Back into the Digital Humanities. Feminism, Generosity, and Mess. In M.K. Gold & L.F. Klein (Eds.), *Debates in the Digital Humanities* 2016 (no pag.). Minneapolis: University of Minnesota Press. URL: https://dhdebates.gc.cuny.edu/read/untitled/section/cfe1b125-6917-4095-9d56-20487aa0b867#ch10 (Accessed: 25 June 2024).

Okuda, N., Kinnison, J., Burns, P., Coffee, N., & Scheirer, W. (2022). Tesserae Intertext Service, *Digital Humanities Quarterly,* 16(1), 1–61. URL: http://www.digitalhumanities.org/dhq/vol/16/1/000602/000602.html (Accessed: 25 June 2024).

Orlandi, T. (2021). Reflections on the Development of Digital Humanities, *Digital Scholarship in the Humanities,* 36(2), 222–229. DOI: https://doi.org/10.1093/llc/fqaa048 (Accessed: 25 June 2024).

Pinter, Y., Marone, M., & Eisenstein, J. (2019). Character Eyes. Seeing Language through Character-Level Taggers. In T. Linzen, G. Chrupala, Y. Belinkov & D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP. Analyzing and Interpreting Neural Networks for NLP* (pp. 95–102). Florenz: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W19-4811 (Accessed: 25 June 2024).

Rawson, K., & Muñoz, T. (2019). Against Cleaning. In M.K. Gold & L.F. Klein (Eds.), *Debates in the Digital Humanities* 2019 (no pag.). URL: https://dhdebates.gc.cuny.edu/read/untitled-f2acf72c-a469-49d8-be35-67f9ac1e3a60/section/07154de9-4903-428e-9c61-7a92a6f22e51#ch23 (Accessed: 25 June 2024).

Schöch, Ch. (2013) Big? Smart? Clean? Messy? Data in the Humanities, *Journal of Digital Humanities,* 2(3), no pag. URL: http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities (Accessed: 25 June 2024).

Schroeder, C.T. (2019). Cultural Heritage Preservation and Canon Formation. What Syriac and Coptic Can Teach Us about the Historiography of the Digital Humanities. In G. Frank, S. Holman & A. Jacobs (Eds.), *The Garb of Being. Embodiment and the Pursuit of Holiness in Late Ancient Christianity* (pp. 318–345). New York: Fordham University Press.

Ead., & Zeldes, A. (2020). A Collaborative Ecosystem for Digital Coptic Studies, *Journal of Data Mining & Digital Humanities,* 1–9. [= *Numéro spécial sur la collecte, la préservation et la diffusion du patrimoine culturel menacé pour de* *nouvelles compréhensions grâce à des approches multilingues*]. DOI: https://doi.org/10.46298/jdmdh.5969 (Accessed: 25 June 2024).

Eid. (2016). Raiders of the Lost Corpus, *Digital Humanities Quarterly,* 10(2), 1–38. URL: http://digitalhumanities.org/dhq/vol/10/2/000247/000247.html (Accessed: 25 June 2024).

Shisha-Halevy, A. (2006). Review of *Coptic Grammar. 2. ed.* by Bentley Layton, *Orientalia,* 75(1), 132–133. URL: https://arielshishahalevy.huji.ac.il/publications2006c (Accessed: 25 June 2024).

Smith, D. A., & Cordell, R. (2018). *A Research Agenda for Historical and Multilingual Optical Character Recognition.* URL: http://hdl.handle.net/2047/D20297452 (Accessed: 25 June 2024).

Wheeler, D., & Jensen, K. (2014). Juxta Commons [poster], *Journal of Digital Humanities,* 3(1), no pag. URL: https://journalofdigitalhumanities.org/3-1/juxta-commons (Accessed: 25 June 2024).

Zeldes, A., & Abrams, M. (2018). The Coptic Universal Dependency Treebank. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (pp. 192–201). Brüssel: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W18-6022 (Accessed: 25 June 2024).

Zeldes, A., Ritz, J., Lüdeling, A., & Chiarcos, Ch. (2009). ANNIS. A Search Tool for Multi-Layer Annotated Corpora. In *Proceedings of Corpus Linguistics 2009.* Liverpool: American Association of Corpus Linguistics. URL: http://ucrel.lancs.ac.uk/publications/cl2009 (Accessed: 25 June 2024).