[*JGRChJ* 9 (2013) 9-41]

# Digitizing Ancient Inscriptions and Manuscripts: Some Thoughts about the Production of Digital Editions

Christopher D. Land
McMaster Divinity College, Hamilton, ON, Canada

## *Introduction*

We live in a world where everything is going digital, so it is not surprising that the digitization of ancient artefacts is becoming a common practice. The question is not *if* ancient artefacts should be digitized, it is *how*. The first section of this article introduces the digital humanities, an emerging field of study that is concerned with the impact of computer technology on human knowledge. A second section then discusses how Greek inscriptions and manuscripts are already being represented using digital technologies. In a third section, I take a step back and reflect on several theoretical distinctions that are relevant to the representation of ancient inscriptions and manuscripts. Then, in a fourth and final section, I discuss why printed editions provide only very selective representations, why recent digital editions seem to be perpetuating this selective focus and how a modular approach might be used to produce more comprehensive editions.

Although I am largely preoccupied with Greek texts, I have chosen not to focus on developments taking place within the field of New Testament studies. My expectation is that readers will already be familiar with such things as the Bibleworks Manuscript Project, the Center for the Study of New Testament Manuscripts (CSNTM), the Codex Sinaiticus Project, the H. Milton Haggard Center for New Testament Textual Studies (HCNTTS), the International Greek New Testament Project (IGNTP), the Institut für Neutestamentliche Textforschung (INTF), the Institute for Textual Scholarship and Electronic Editing (ITSEE), the Online Critical Pseudepigrapha and the like. My failure to discuss these efforts does not reflect their unimportance; rather, it reflects my desire to engage with the digital humanities more broadly.

## *The Digital Humanities*

The digital humanities, also known as humanities computing, is an interdisciplinary field of study concerned with the relationships that exist between human knowledge and digital media.[1] Because these relationships are bidirectional, the field of digital humanities does not merely study how scholars employ computational tools; it also studies how computational tools affect the scholarship of those who use them.

A basic principle that must be kept in view is this: the advent of new technologies is not merely an opportunity to do old things in new ways, but an invitation to re-consider what might be done. As Roberto Busa observes, ‘The use of computers in the humanities has as its principal aim the enhancement of the quality, depth and extension of research and not merely the lessening of human effort and time’.[2] This point is so essential to the spirit of the digital humanities that I must quote Busa’s comments at greater length:

> In this field one should not use the computer primarily for speeding up the operation, nor for minimizing the work of the researchers. It would not be reasonable to use the computer just to obtain the same results as before, having the same qualities as before, but more rapidly and with less human effort… Today’s academic life seems to be more in favor of many short-term research projects which need to be published quickly, rather than of projects requiring teams of co-workers collaborating for decades… [But] it would be much better to build up results one centimetre at a time on a

[1] An overview of the digital humanities is available in Susan Schreibman, Ray Siemens and John Unsworth (eds.), *A Companion to Digital Humanities* (Oxford: Blackwell, 2004). Online: http://www.digitalhumanities.org/companion. Some issues involved in defining the digital humanities are confronted in Patrik Svensson, ‘The Landscape of Digital Humanities’, *Digital Humanities Quarterly* 4.1 (2010), n.p. Online: http://www.digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

[2] Roberto Busa, ‘The Annals of Humanities Computing: The Index Thomisticus’, *Computers and the Humanities* 14 (1980), pp. 83-90 (89). Busa is undeniably the father of humanities computing. In 1949, he began to create an index of all the words in the works of Thomas Aquinas and several related authors, a project that eventually totalled some 11 million Latin words. Deciding that machines might assist him in his efforts, he enlisted the support of Thomas Watson at IBM, and eventually created the first lemmatized digital concordance. For an account of these developments, see Susan Hockey, ‘The History of Humanities Computing’, in Schreibman, Siemens and Unsworth (eds.), *A Companion to Digital Humanities*, pp. 3-19.

> base one kilometre wide, than to build up a kilometre of research on a one-centimetre base.[3]

The essential idea that Busa communicates in this passage is his conviction that the process of going digital has the potential to *transform* scholarship at many levels. It not only changes how scholars encounter their data; it changes how they conceive of analysis. It not only changes how work is disseminated to other scholars; it opens up new possibilities for collaborative effort. It is not merely scholarly resources that must be digitized; to a certain extent, scholarship itself must be digitized.[4]

Among other things, this process entails the establishment of online communities and online research tools. Scholars working in the digital humanities model this quite well. Many of their resources are situated on the internet, including some that are traditional (e.g. text resources) and some that are novel (e.g. wikis, blogs, discussion groups, podcasts, version control systems, etc.). Of particular interest is The Digital Classicist, a web-based hub that is hosted by the Centre for Computing in the Humanities at King's College London.[5] The Digital Classicist serves as an access point for people who wish to connect with the digital humanities, making it easy to get in touch with other scholars, to discover projects that are being undertaken, to find out about events that are taking place around the world and to locate useful resources. It also contains a rapidly growing wiki that facilitates cooperative research and cooperative learning.[6] For biblical scholars who are interested in

[3] Busa, 'Annals of Humanities Computing', p. 89.

[4] In a recent article about the relationship between the field of classics and computer technology, Greg Crane makes the fascinating point that 'While many [non-classicists] compare the impact of print and of new electronic media, classicists can see the impact of both revolutions upon the 2,500-year history of their field' ('Classics and the Computer: An End of the History', in Schreibman, Siemens and Unsworth [eds.], *A Companion to Digital Humanities*, pp. 46-55 [46]). Crane's article should be read alongside Theodore Brunner's earlier article ('Classics and the Computer: The History of a Relationship', in J. Solomon [ed.], *Accessing Antiquity: The Computerization of Classical Studies* [Tucson: University of Arizona Press, 1993], pp. 10-33), since together they provide a historical overview of computing within the field of classics.

[5] 'The Digital Classicist: Advanced Digital Methods Applied to the Study of the Ancient World', n.p. Online: http://www.digitalclassicist.org. It should be noted that the hub is not funded or owned by any institution, but is operated by a decentralized and collaborative community.

[6] For more on the Digital Classicist wiki, see Simon Mahony, 'Research

exploring the future (for a change), The Digital Classicist is a great place to start. It demonstrates that an online network can attract new people to a field, while simultaneously bringing together scholars who are already well established in their field. Another good place to stay up-to-date is the blog of the Stoa Consortium, which supplies news and commentary concerning digital applications, methodology and resources.[7]

The idea of an online community of scholars is hardly revolutionary. However, the process of digitization entails two additional developments: a prioritizing of collaborative efforts and a move towards the open distribution of research data and research publications.[8] Once again, examples from the digital humanities are encouraging. The Suda On Line is in the process of translating and annotating the entire Suda, a tenth-century Byzantine encyclopaedia. To date, over 170 scholars have participated and 25,000 entries have been completed.[9] The Homer Multitext Project seeks to present the transmission history of Homer's *Iliad* and *Odyssey* in an open source format.[10] It is making available a vast library of machine-readable texts and images, along with indices and software tools that will enable scholars to interact with those texts and images. Currently, over thirty scholars are affiliated with the project.

With regard to the open distribution of research results, positive examples are emerging with increasing frequency. For instance, the peer-reviewed journal *Digital Humanities Quarterly* prides itself on being an open-access publication, employs open standards to deliver its content

Communities and Open Collaboration: The Example of the Digital Classicist Wiki', *Digital Medievalist* 6 (2011), n.p. Online: http://www.digitalmedievalist.org/journal/6/mahony/.

[7] 'The Stoa Consortium', n.p. Online: http://www.stoa.org.

[8] For various perspectives on the topic of cyberinfrastructure, see the articles in *Digital Humanities Quarterly* 3.1 (2009).

[9] Suda On Line and the Stoa Consortium, 'About the Suda On-Line', n.p. Online: http://www.stoa.org/sol. For more information about the Suda On Line, see Anne Mahoney, 'Tachypaedia Byzantina: The Suda On Line as Collaborative Encyclopedia', *Digital Humanities Quarterly* 3.1 (2009), n.p.

[10] For an overview of the project, see Gregory Nagy, 'The Homer Multitext Project', in Jerome McGann, Andrew M. Stauffer and Dana Wheeles (eds.), *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* (Houston: Rice University Press, 2010), pp. 87-112. Online: http://cnx.org/content/col11199/1.1. For the latest news, see the Homer Multitext blog. Online: http://homermultitext.blogspot.com.

and even publishes a blog with guest commentators. The *Bulletin of the American Society of Papyrologists* is now entirely online and available for free. And the Inscriptions of Aphrodisias Project has published online all of its texts, translations and commentaries under a Creative Commons licence. Hopefully, these initiatives are indicators of a forward-reaching trend that will eventually make the open distribution of research data and analysis a widespread practice.

Two scholars from the digital humanities, Simon Mahony and Gabriel Bodard, describe the ethos of their discipline as its 'most striking and successful' feature. They write:

> Digital Classicists do not work in isolation; they develop projects in tandem with colleagues in other humanities disciplines or with experts in technical fields: engineers, computer scientists and civil engineers. They do not publish expensive monographs destined to be checked out of libraries once every few years; they collect data, conduct research, develop tools and resources, and importantly make them available electronically, often under free and open licences such as Creative Commons, for reference and for re-use by scholars, students and non-specialists alike.[11]

Unfortunately this way of doing scholarship is, as Dickie Selfe has pointed out, 'a challenge to our own academic cultures'.[12] The *Report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences* states:

> Despite the demonstrated value of collaboration in the sciences, there are relatively few formal digital communities and relatively few institutional platforms for online collaboration in the humanities. In these disciplines, single-author work continues to dominate. Lone scholars, the report remarked, are working in relative isolation, building their own content and tools, struggling with their own intellectual property issues, and creating their own archiving solutions.[13]

[11] Gabriel Bodard and Simon Mahony, 'Introduction', in Gabriel Bodard and Simon Mahony (eds.), *Digital Research in the Study of Classical Antiquity* (London: Ashgate, 2010), pp. 1-11 (2).

[12] As cited in 'American Council of Learned Societies, Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences', (2006), p. 21. Online: http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf.

[13] 'American Council of Learned Societies, Our Cultural Commonwealth', p. 21.

Computer technology has been making headway into the humanities for quite some time, and attitudes towards online collaboration and publication have already softened. It is my fervent hope that the field of biblical studies will continue to support these general developments, particularly with respect to the digitization and distribution of ancient texts. The immense labour involved in publishing new primary texts should not prohibit the open distribution of those publications. Rather, it should motivate scholars to pursue collaborative work and to be grateful when new publications are made freely available. Similarly, the prestige of those journals that require paid subscriptions should not cause us to disregard more open forms of distribution. Rather, we should seek to bolster the reputation of those open-access journals that publish respectable, peer-reviewed research.

## *Digital Resources: The Current State of Affairs*

### *Digital Repositories*

Digital resources have been available to classicists, epigraphers, papyrologists and biblical scholars for quite some time, although some of these resources have undergone significant changes in recent years. The largest and best-known database of ancient Greek texts is the Thesaurus Linguae Graecae (TLG). Established in 1972, the TLG presents itself as the first major application of computer technology to the discipline of classical scholarship. Its goal is to create a comprehensive database of Greek literature, and it contains more than 105 million words from over 10,000 works associated with 4,000 authors. As the guinea pig of digital classicism, the TLG has undergone some dramatic revisions over the course of nearly four decades. Most notably, it has transitioned from being a CD-ROM-based Beta Code resource to being a web-based Unicode resource, and it has progressed from being a simple corpus to being a lemmatized corpus. Although the TLG maintains a proprietary stance towards all of its digital materials, the Perseus Project distributes its texts under a Creative Commons licence. Perseus contains a smaller digital collection with just over 8 million words of literary Greek. It actively supports open-source initiatives, has published its source code online and even distributes XML editions of its contents.[14] The Perseus

[14] Greg Crane, 'Plutarch, Athenaeus, Elegy and Iambus, the Greek Anthology, Lucian and the Scaife Digital Library—1.6 Million Words of Open Content Greek', n.p. Online: http://www.stoa.org/archives/1332.

Project is currently creating a syntactic database of classical Greek.[15]

Because both the TLG and Perseus are literary collections, they are not really concerned with representing individual manuscripts. For this reason other collections are more relevant here. The Duke Databank of Documentary Papyri (DDBDP) is an electronic corpus of both Greek and Latin documents. It began in 1982 and now contains nearly 500 papyrus volumes. These documents used to be hosted by Perseus, but they have recently been integrated into the Papyrological Navigator (see below). A complementary project, the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV), began in 1988. It seeks to produce a comprehensive repository of images and metadata relating to all published Greek or Latin documentary papyri, and it has made over 56,000 records of this nature freely available online.[16] The metadata associated with each document records its title, date, provenance and writing material, as well as providing bibliographic information. Wherever possible, images and translations are also supplied. A similar but more broadly focused endeavour, the Advanced Papyrological Information System (APIS), went online in 1995 as a web-based databank of images and metadata pertaining to written material in a variety of ancient languages. The APIS databank now possesses over 33,000 records.[17]

In recent years, exciting developments have taken place involving all three of these resources. In 2004–2005, the DDBDP and HGV began the task of mapping their complementary data-sets to one another. Soon thereafter, funding was received for Integrating Digital Papyrology (IDP1), a collaborative effort designed to establish a sustainable future for the resources contained in DDBDP and to help the field of papyrology pursue the interoperation of its digital resources.[18] As a result of this

[15] David Bamman, Francesco Mambrini and Gregory Crane, 'An Ownership Model of Annotation: The Ancient Greek Dependency Treebank', in *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT-8)* (Milan: Northern European Association for Language Technology [NEALT], 2009), pp. 5-15. Online: http://hdl.handle.net/10427/70399.

[16] Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV), 'Einführung', n.p. Online: http://www.rzuser.uni-heidelberg.de/~gv0/Einfuehrung.html.

[17] Columbia University Libraries Digital Program, 'APIS Central Database: Selected Statistics', n.p. Online: http://www.columbia.edu/cu/libraries/inside/projects/apis/statistics.

[18] Roger Bagnall, 'Integrating Digital Papyrology' (paper presented at Online Humanities Scholarship: The Shape of Things to Come, University of Virginia, 26-28

project and its successors IDP2 and IDP3, the texts of the DDBDP and the data in HGV both have been migrated to XML in conformance with the internationally recognized EpiDoc standard (see below), and the results have been integrated with the APIS metadata through a single online search-interface entitled the Papyrological Navigator.[19]

The Papyrological Navigator provides its users with several routes into its corpus. Most importantly, papyri may be retrieved through a search interface that permits the use of multiple criteria (e.g. previous publication, provenance, language, date, etc.). As an additional benefit, the Navigator permits a full-text lemmatized search of all the papyri made available through the DDBDP. Once a papyrus has been chosen for viewing, the user is taken to an overview page that contains sections dedicated to the following: HGV metadata, APIS metadata, the DDBDP transcription, translation(s) and images. Because the information supplied by the Navigator is retrieved from various existing resources, the layout is somewhat less elegant than would be possible for a fresh publication, but the convenience of having everything in one place is a significant advance. What is more, a web-based editing platform entitled the Papyrological Editor has been developed that enables scholars from around the world to add texts to the DDBDP, correct typos, add or change translations and propose emendations. Because accessibility is essential for scholarly collaboration, the Editor uses a shorthand called Leiden+, which closely resembles the Leiden conventions.[20]

The Catalogue of Paraliterary Papyri (CPP) should not go unnoticed, since it provides digital versions of many papyri that are not online elsewhere.[21] Nor should a number of very important epigraphical databases, especially those that are part of the Electronic Archive of Greek and Latin Epigraphy (EAGLE).[22] The Packard Humanities Institute maintains an archive of Greek epigraphical texts.[23] Trismegistos is also

March 2010). Online: http://shapeofthings.org/papers.

[19] NYU Digital Library Technology Services and the Institute for the Study of the Ancient World, ‘Papyri.info’, n.p. Online: http://papyri.info.

[20] Joshua Sosin, ‘Digital Papyrology’, paper presented at the Congress of the International Association of Papyrologists, Geneva, Switzerland, 19 August 2010. Online: http://www.stoa.org/archives/1263.

[21] K.U. Leuven Research Unit Greek Studies, ‘Catalogue of Paraliterary Papyri (CPP)’, n.p. Online: http://cpp.arts.kuleuven.be.

[22] Federazione Internazionale di Banche dati Epigrafiche, ‘Electronic Archive of Greek and Latin Epigraphy’, n.p. Online: http://www.eagle-eagle.it.

[23] The Packard Humanities Institute, ‘Searchable Greek Inscriptions’, n.p.

a valuable resource, providing papyrological and epigraphical resources dealing with Egypt and the Nile valley roughly between 800 BCE and 800 CE.[24]

*Digital Imaging Technologies*

In recent years, many older facsimiles and photographic plates have been digitized. Moreover, new photographs have been taken using improved photographic technologies. But perhaps most exciting of all is the way that new imaging technologies are actually enhancing our ability to read ancient manuscripts. Pride of place in this respect must go to multi-spectral imaging, a technique that captures images at various wavelengths, including many that are not visible to the human eye. This technology will noticeably increase the amount of information that is available to editors of papyri. Researchers at Oxford University, where a digital imaging process is being applied to the Oxyrhynchus collection, have suggested that the technique could increase our collection of ancient writing by about twenty percent.[25]

Multi-spectral imaging not only captures information that would be otherwise inaccessible to human observers, it also permits researchers to isolate specific electromagnetic wavelengths. Because different materials reflect light differently, documents that are completely illegible when viewed normally may become perfectly clear when certain spectra are isolated. The team at Oxford University has published an online demonstration of how this works, using P.Oxy. XXX 2507 as an example.[26] Technologists can also bring several wavelengths together in order to form composite images. A dramatic example of this may be seen in the recent Archimedes Palimpsest Project. The Project centres on a thirteenth-century prayer book containing several erased texts, including two previously unknown treatises by Archimedes. A wide range of spectral images was produced of the book, involving twelve wavelengths of LED illumination (some applied as raking light), as well as UV and

Online: http://epigraphy.packhum.org.

[24] 'Trismegistos: An Interdisciplinary Portal of Papyrological and Epigraphical Resources Dealing with Egypt and the Nile Valley between Roughly 800 BC and AD 800', n.p. Online: http://www.trismegistos.org/index.html.

[25] James Owen, 'Papyrus Reveals New Clues to Ancient World', *National Geographic News* (2005), n.p. Online: http://news.nationalgeographic.com/news/2005/04/0425_050425_papyrus.html.

[26] The Imaging Papyri Project, University of Oxford, 'Multispectral Imaging', n.p. Online: http://www.papyrology.ox.ac.uk/POxy/multi/index.html.

tungsten illumination. X-ray images were also produced, using front and back detectors with channels at various energies. Special processing techniques were then used to integrate the raw spectral images into four different display images, including one set that has been published online by Google.[27] A side-by-side comparison of a composite image and a normal photographic image reveals the great potential of multi-spectral imaging. The erased Archimedes texts, which are barely visible beneath the current text of the prayer book when it is viewed under normal conditions, come into sharp relief when imaged using multi-spectral technology.

Of course, in working with a codex in relatively good condition, the Archimedes Palimpsest team had a much better data source than is often available. Many ancient scrolls are not simply illegible—they cannot be opened. Here, too, modern imaging techniques are paving the way forward. The EDUCE Project (Enhanced Digital Unwrapping for Conservation and Exploration) is working to develop a hardware and software system for the virtual unwrapping and visualization of ancient texts. Their technique involves the use of a custom-built, portable, multi-power CT scanning device that simultaneously sends out X-rays from different angles. Once a ‘slice’ of a scroll has been captured, the mechanism shifts position by a microscopic amount and then scans again. Eventually, the data is fed into a software program that forms a complete cross-sectional image and then digitally unrolls the scroll.[28] Early images captured using the technique suggest that it does indeed have the potential to expand our collection of available resources.[29] Hopefully it will allow us to image texts that are currently inaccessible. At the very least, it will facilitate the imaging of warped or distorted documents.

Alongside the production of new digital images, we must also consider their distribution. After all, the move from photographic plates to computer files makes it much easier for scholars to exchange their

[27] The Archimedes Palimpsest Project, ‘The Archimedes Palimpsest’, n.p. Online:http://archimedespalimpsest.org/digital/google-book.php.

[28] Alicia P. Gregory, ‘Digital Exploration: Unwrapping the Secrets of Damaged Manuscripts’, *Odyssey* (2004), pp. 18-23. Online: http://www.research.uky.edu/odyssey/fall04/seales.html.

[29] Ryan Baumann, Dorothy Carr Porter and W. Brent Seales, ‘The Use of Micro-CT in the Study of Archaeological Artifacts’, paper presented at the Ninth International Conference on NDT of Art, Jerusalem, Israel, 25-30 May 2008. Online: http://212.8.206.21/article/art2008/papers/244Seales.pdf.

primary resources. Presently, there is no centralized repository that stores, organizes and distributes new manuscript images. Rather, images are managed by specific institutions or projects. This is a workable arrangement, provided that the images in question are openly distributed. A very positive precedent in this respect has been set by the Archimedes Palimpsest Project, which has published all of its raw imaging data online under a Creative Commons licence.

The Archimedes Palimpsest Project is also illustrative of yet another important development. As part of the Project, Optical Character Recognition (OCR) software is being developed that will not only transcribe visible characters but will facilitate the reconstruction of partial characters.[30] Using traditional OCR technology, a computer determines the probabilities that a certain marking is an instance of a certain Greek character. These probabilities are then refined by looking at detailed spatial characteristics and by considering the likelihood of each possible character occurring before or after any neighbouring characters. As the project website clearly states, 'The results will facilitate scholars by presenting them with a range of possibilities from which they might choose'.[31] What this means, in practical terms, is that the production and distribution of high quality multi-spectral images will soon lead into a semi-automated process of digital transcription. It follows that encoding standards must be developed in such a way that they can easily interface with OCR transcription technology on the one hand, and yet effectively accommodate subsequent editorial tasks on the other.

### *Digital Encoding Standards*

The development of encoding standards for the digitization of primary texts is a matter of considerable significance to the digital humanities. But of course, when humanities scholarship first began to exploit the advantages of digital technology, sustainability and interoperability were not the pressing issues they are today. Many early systems were

[30] See also the work discussed in Arianna Ciula, 'The Palaeographical Method under the Light of a Digital Approach', in Malte Rehbein, Patrick Sahle and Torsten Schaßan (eds.), *Kodikologie und Paläographie im digitalen Zeitalter—Codicology and Palaeography in the Digital Age* (Schriften des Instituts für Dokumentologie und Editorik, 2; Norderstedt: Books on Demand [BoD], 2009), pp. 219-35. Online: http://kups.ub.uni-koeln.de/2971.

[31] The Archimedes Palimpsest Project, 'Optical Character Recognition', n.p. Online: http://archimedespalimpsest.org/about/imaging/optical-character-recognition.php.

poorly designed and quickly became obsolete, and they were rarely compatible with one another. In hindsight, it is clear that this situation '[inhibited] the development of the full potential of computers to support humanistic inquiry by erecting barriers to access, creating new problems for preservation, making the sharing of data (and theories) difficult, and making the development of common tools impractical'.[32] Eventually, at Vassar College in November 1987, a number of scholars and other academic professionals gathered together to address what was by then a rapidly growing problem. Out of this meeting emerged the Text Encoding Initiative (TEI), which is now a member-funded non-profit corporation. TEI aims 'to develop, maintain, and promulgate hardware- and software-independent methods for encoding humanities data in electronic form'.[33] The original draft Guidelines from the project, as well as the first official Guidelines, used Standard Generalized Markup Language (SGML); since 2002, however, the TEI Guidelines have provided the digital humanities with an Extensible Markup Language (XML) standard for the representation of texts in digital form. This standard is now widely acknowledged and implemented. The National Endowment for the Humanities, for instance, states that applicants for its Scholarly Editions and Translations grants 'are encouraged to use open standards and markup conforming to the Text Encoding Initiative (TEI), and to employ current best practices in the creation of electronic editions'.[34]

The TEI Guidelines provide two key resources: first, a modular, extensible XML scheme; and secondly, detailed documentation that explains important concepts, describes proper usage and exemplifies how the tagging works. The XML scheme works like any other in that it defines a set of elements, along with attributes that modify those elements. Because the Guidelines are designed to facilitate the encoding of any possible text, they are extremely robust: the full tag set consists of roughly 500 different elements. The TEI scheme is modular in the sense that it distinguishes a small core set of tags from all the others, which means that users can adopt and combine the additional tags in whatever way best suits their needs. The scheme is extensible in that it permits users to add, redefine or rename elements as necessary, and provides

[32] TEI Consortium, 'TEI: History', n.p. Online: http://www.tei-c.org/About/history.xml.

[33] TEI Consortium, 'TEI: History.'

[34] National Endowment for the Humanities, 'Scholarly Editions and Translations', n.p. Online: http://www.neh.gov/grants/guidelines/editions.html.

specific procedures for how this ought to be done.

One customization of TEI that will be of particular interest to classicists, epigraphers, papyrologists and biblical scholars is the EpiDoc standard. EpiDoc was launched as a public enterprise in the late 1990s in response to a manifesto by Silvio Panciera that called for a free and unrestricted database of all surviving Greek and Latin epigraphical texts. The original author of the EpiDoc standard, a graduate student named Tom Elliott, envisioned that XML would provide classical epigraphists with a means of digitizing the editorial conventions that had been developed early in the twentieth century (i.e. the Leiden conventions). Hence, the EpiDoc website states, 'EpiDoc must facilitate the encoding of all editorial observations and distinctions signaled in traditional print editions through the use of sigla and typographic indicia'.[35] Stylesheets are available that will display any EpiDoc text in a human-readable version that adheres to the Leiden conventions. What is more, a software tool entitled the Chapel Hill Electronic Text Converter (CHETC) can convert any text using standard typographic conventions into EpiDoc-compliant XML. Because of its emphasis on backwards compatibility, EpiDoc is a very useful standard.

Once an EpiDoc edition is complete, there are various ways that it can be displayed to end-users. A good example of this may be seen in the online *Inscriptions of Aphrodisias* (InsAph).[36] The InsAph website offers its users more entry points into its corpus than are generally provided by traditional print publications. It is possible to link through a series of tables that list inscriptions according to their text-type, monument-type, decorative features and date. An interactive map of Aphrodisias permits the retrieval of inscriptions according to their location, and, of course, users are able to search for inscriptions using a range of criteria. Once a specific inscription has been selected for viewing, the InsAph website loads all the information relevant to that inscription on a single page.

[35] Tom Elliott, 'EpiDoc: Epigraphic Documents in TEI XML', n.p. Online: http://epidoc.sourceforge.net/index.shtml.

[36] Charlotte Roueché, *Aphrodisias in Late Antiquity: The Late Roman and Byzantine Inscriptions* (rev. 2nd edn, 2004), n.p. Online: http://insaph.kcl.ac.uk/ala2004; Gabriel Bodard, 'The Inscriptions of Aphrodisias as Electronic Publication: A User's Perspective and a Proposed Paradigm', *Digital Medievalist* 4 (2008), n.p. Online: http://www.digitalmedievalist.org/journal/4/bodard; Charlotte Roueché, 'Digitizing Inscribed Texts', in Marilyn Deegan and Kathryn Sutherland (eds.), *Open Source Critical Editions: A Rationale* (Farnham, UK: Ashgate, 2009), pp. 159-68.

At the top, the title of the inscription is given, along with a description of the physical object on which it is inscribed, a description of its text, a description of its letters and a statement concerning its dating. Next, a brief history of the object is given, listing first the relevant locations (findspot, original location, last recorded location) and then a prose narrative of its discovery and subsequent treatment. A traditional bibliography follows and then the primary text itself. It is here that the advantages of XML come through most clearly. The user is given a choice between three modes of display: 'Edition', 'Diplomatic' and 'EpiDoc'. The first mode follows the Leiden conventions; the second displays majuscule characters with no accents or spacing and offers no reconstructions; the third displays the underlying XML document that is being used to generate the other two. Beneath the primary text, an apparatus is provided, then a translation, additional commentary and finally any available photographs of the inscription. In addition to the edited inscriptions, the InsAph website provides a number of indices, listing all Greek and Latin lemmata, fragmentary words, names (of both persons and places), special characters, ligatured characters, numerals, abbreviations and expansions.

## *Making an Edition: Some Guiding Principles*

With the advent of new imaging technologies and encoding languages it is becoming possible to publish inscriptions and manuscripts in entirely new ways. However, implementations of these technologies need to be guided by sound principles.[37] I suggest that the following factors need to be considered in the publication of ancient inscriptions and manuscripts, irrespective of whether a print or digital medium is in view: (1) levels of abstraction; (2) modes of representation; (3) domains of reconstruction (see Table 1).

[37] Discussing the move to digital texts is hardly a new exercise. For a good entry into the relevant literature, I recommend the following issues devoted to the creation of digital editions: *Literature Compass* 7.2 (2010) (Special Issue: *Scholarly Editing in the Twenty-First Century*); *Literary and Linguistic Computing* 24 (2009) (Special Issue: *Computing the Edition*). Numerous references to the wider literature can be found in the articles contained therein.

Table 1: Levels of Abstraction and their Related Objects

| **Levels of Abstraction** | | **Objects** | | **Modes of Representation** | **Default Representations** | **Domains of Reconstruction** |
|---|---|---|---|---|---|---|
| ARCHAEOLOGY | | ARTEFACTS | | Visual | Images | |
| PALAEOGRAPHY & GRAPHETICS | | SYMBOLS | | Symbolic | Graphetic Transcriptions | Physical Deterioration over Time |
| | Graphology | Writing | TEXTS | Linguistic | Graphemic Transcriptions | Scribal Error in Writing |
| | Lexicogrammar | Wording | | | Rewrites | Scribal Error in Wording |
| | Semantics | Meaning | | | Translations | |

*Levels of Abstraction*

With regard to the objects that they observe and study, scholars of antiquity are often divided. They are torn between the study of archaeological *artefacts* and the study of ancient *texts*. Or to put things differently: they want to study both matter and meaning. Take papyrology, for example. The physical objects studied by papyrology have a material composition (e.g. papyrus, ink), a material form and appearance (e.g. size, shape, colour) and also a material history (e.g. age, place of origin, state of deterioration, place of discovery, current location). By way of contrast, the linguistic objects studied by papyrology are only discernable when viewed in the light of abstract systems of language. But, of course, physical artefacts and linguistic texts are not really separate objects of study. Rather, the papyrologist who studies an ancient piece of papyrus does so at different levels of abstraction, such that linguistic realities are dependent upon concrete physical realities. Wherever deterioration has affected the writing on a piece of papyrus, part of an artefact has gone missing—but also part of a text.

Looking a little more closely at this situation, it becomes apparent that there is no simple dichotomy between artefacts and texts. In reality, there are a number of different levels of abstraction as one's attention turns gradually from a material manuscript towards a meaningful text. In this article, my focus is mostly on non-physical levels of analysis. The least abstract of these is called graphetics, by analogy with the more familiar discipline of phonetics. Like palaeography, it is concerned with the visual appearance of symbolic markings and with their spatial arrangement on writing surfaces. Moving progressively higher in abstraction and closer to the notion of meaning, we pass through graphology (which is akin to phonology), and then lexicogrammar (which encompasses the traditional domains of morphology, lexis and syntax), until we finally reach semantics (see Table 1).[38] Each of these levels of abstraction defines its own objects, although more abstract objects are manifested through less abstract objects.

### *Modes of Representation*

In addition to examining multiple objects, fields like epigraphy and papyrology employ several different modes of representation. A representation is created whenever some object is transformed by an observer into information and then re-presented in a new form. Because there are so many ways of communicating information, there is an almost limitless number of ways to represent an inscription or manuscript. Here I will discuss what I take to be the three default modes of representation that are relevant to the making of a scholarly edition: the visual mode, the symbolic mode and the linguistic mode. The first of these uses imagery to represent the appearance of a physical artefact. The second uses symbols to represent the symbolic markings on a physical artefact. The third uses writing to represent the writing, wording and meaning of an ancient text. These are not, of course, the only possible modes of representation that might be employed in a scholarly edition. But they are the default ones. In what follows, I will show how these three modes can be used to create five 'default' representations, each of which is oriented towards a specific level of abstraction.

*Images* have traditionally been photographic, although recent technologies are producing other sorts of images. The essential thing about imaging is that it creates a representation of a physical object using some form of electromagnetic radiation and some sort of sensor. Like all

[38] It must be stressed that within this framework the terms graphology and graphetics are analogous to phonology and phonetics. Accordingly, graphology is not being used to signify the psychological study of handwriting. Graphetics may be subdivided into visual and mechanical graphetics by analogy with auditory and articulatory phonetics. Orthography is a prescriptive discipline that applies standard rules to the production of written texts. Typography is a subdivision of graphology that deals with printed texts. For a helpful treatment of these and other related terms, see Florian Coulmas, *The Blackwell Encyclopedia of Writing Systems* (Oxford: Blackwell, 1996) and David Crystal, *A Dictionary of Linguistics and Phonetics* (Oxford: Blackwell, 6th edn, 2008).

representations, images are selective. Their selection can be very broad, as when the entire spectrum of visible light is captured, or very narrow, as when specific spectra are isolated in order to bring out more clearly the markings on a document.

Selection is a necessary part of representing symbols as well. The difference here is that a higher degree of abstraction is involved, since one must interpret physical markings as tokens of specific symbol types. This means excluding from consideration a number of markings that are regularly found on ancient artefacts, such as ink blotches, decorations and scoring lines, while taking into account such things as letters, diacritics, punctuation marks, canon numbers and musical notations. It also means that when two distinct physical tokens are classified as instances of a single type and are represented as such, any additional information contained in the original tokens is filtered out and lost.

As a bare minimum, any textually oriented representation of an inscription or manuscript must account for those symbols that are linguistic in nature. However, linguistic symbols can be accounted for in various ways. An editor might represent the distinctive *form* of writing employed by a scribe; alternatively, the editor might represent merely the *content* of what is written. Although in both cases the resulting representations are referred to as *transcriptions*, it is possible to systematically distinguish between two different types of transcription using the levels of abstraction that were introduced above.[39]

The basic unit of graphology is the grapheme, although some languages possess larger graphological units (often delimited using spaces, punctuation marks, paragraph markers, etc.). Speaking linguistically, a grapheme is the minimal contrastive unit in the writing system of a language. This means that graphemes have the potential to differentiate between different lexicogrammatical forms (e.g. the mu and sigma in

[39] Concerning some possible types of transcription, see Peter Robinson and Elizabeth Solopova, ‘Guidelines for Transcription of the Manuscripts of the Wife of Bath’s Prologue’, in Norman Blake and Peter Robinson (eds.), *The Canterbury Tales Project Occasional Papers I* (Oxford: Office for Humanities Communication, 1993), pp. 19-52. Online: http://www.canterburytalesproject.org/pubs/transguide.pdf; Dominique Stuzmann, ‘Paléographie statistique pour décrire, identifier, dater: Normaliser pour coopérer et aller plus loin?’, in Franz Fischer, Christiane Fritze and Georg Vogeler (eds.), *Kodikologie und Paläographie im digitalen Zeitalter 2—Codicology and Palaeography in the Digital Age 2* (Schriften des Instituts für Dokumentologie und Editorik, 3; Norderstedt: Books on Demand [BoD], 2011), pp. 247-77. Online: http://kups.ub.uni-koeln.de/id/eprint/4353.

μου and σου). Graphetics, on the other hand, is concerned with graphs, which may be defined as the smallest discrete segments in a stretch of writing. For the most part, graphs can be thought of as different ways of expressing the graphemes of a language, whether individually (e.g. a supralinear stroke used for nu at the end of a line) or in combination (e.g. ligatures). Applying these distinctions to the topic at hand, it can be said that a *graphetic transcription* will seek to represent the distinct graphs that are found on an ancient manuscript. A *graphemic transcription*, on the other hand, will restrict itself to representing graphemes. The former records how symbolic markings are arranged spatially. The latter records how writing is laid out in columns, lines, verses, etc.

While the above distinction may seem like nitpicking, it actually exemplifies a very important generalization: a linguistic representation of a higher-order linguistic abstraction will treat lower-order distinctions as irrelevant detail. Thus a graphological transcription will fail to distinguish between distinct graphs, just as a graphetic transcription will fail to capture all of the detail that would be visible in a photograph. If it were otherwise, there would be no point in designing a different representation for each level of abstraction: every graphological transcription would also be a graphetic transcription, and every graphetic transcription would also be a photographic image—and hence, we would have only images. As things stand, the usefulness of a graphetic transcription derives precisely from the fact that crisp typeset graphs are introduced *in place of* faded handwritten graphs. Similarly, the usefulness of a graphological transcription derives from the fact that modern graphs are introduced *in place of* potentially unfamiliar ancient graphs.

It follows from this way of looking at things that it is possible to define a default representation that captures the wording of a manuscript but overlooks the details of its writing. Lacking a suitable term for such lexicogrammatical representations, I have opted to call them *rewrites*.[40] This seems an appropriate label, given that the purpose of these representations is to manifest the lexicogrammatical content of an ancient text by writing it out afresh. The advantage of creating a rewrite rather than a transcription is the fact that one can introduce modern graphological conventions such as word divisions, diacritics, punctuation

[40] Robinson and Solopova use the phrase *regularized transcriptions*, but they conceive of such transcriptions only in terms of regularized spelling ('Guidelines for Transcription', p. 22). The phrase thus suits their purposes, but it does not distinguish between writing and wording.

and mixed case. One can structure a text using a modern page layout, and one can incorporate scribal corrections into the main text as variant readings. To do such things in a transcription would be to *misrepresent* the writing of an ancient scribe. In the context of a rewrite, however, they serve to *effectively represent* the wording of an inscription or manuscript.

Taking one final step, we reach *translations*. A translation, after all, is nothing more than a semantic representation that disregards lower levels of abstraction, including even the grammar and lexis of the language in which a text was originally composed. As with all of the other representations, here too the goal is to make an ancient object clearer by presenting it in a new form. In this case, that form happens to be a modern language.

### *Domains of Reconstruction*

Once a specific object and representation have been selected, another important distinction arises: an editor must choose whether to rely strictly on *observation* or whether to allow a measure of *reconstruction*. Whereas observation is focused on an *actual* object, reconstruction is focused on an *ideal* object.

In the case of imaging, it is relatively easy to distinguish between actual objects and ideal objects. A physical manuscript as it exists today is an actual object, as can be seen from the fact that light bounces off it before reaching an imaging sensor. That same physical manuscript projected back to some moment in the past—perhaps before its appearance was changed in some significant way, whether through damage or scribal erasure—is an ideal object. Because one cannot bounce light off an ideal object, images cannot be reconstructive. Handwritten facsimiles or digitally edited files might be used to visually reconstruct a damaged inscription or manuscript, but the usefulness of such reconstructions would be limited. Instead, the burden of physical reconstruction is borne primarily by graphetic transcriptions.

The symbolic markings that are the focus of graphetic transcriptions exist at a fairly low level of abstraction, so they are immediately affected by physical changes to an inscription or manuscript. Nevertheless, they are sufficiently abstract that the process of creating a reconstructive representation is usually quite feasible. For this reason, whenever an editor perceives that the physical appearance of an inscription or manuscript has changed over time, whether through damage or through some deliberate process, he must decide whether to represent the graphs that are presently visible or the graphs that he thinks were previously visible. In practice this is no simple black and white choice, since individual graphs may be more or less legible. As a general rule, the degree of idealization that is involved in representing a graph gradually increases as the physical condition of that graph deteriorates. Moreover, as the physical evidence for a graph diminishes, the reconstructive process must rely increasingly on linguistic abstractions that have been obtained from surrounding graphs. Eventually, when a graph is no longer visible at all, reconstruction becomes a downward process entirely dependent on higher-order abstractions (i.e. linguistic context).[41] So then, a reconstructive graphetic transcription is concerned with *physical changes* and hence a *physically* ideal object, even though it relies upon linguistic abstractions.

An interesting but not very surprising thing happens as one begins to represent more abstract linguistic objects such as graphemes and words: the physically ideal artefact becomes completely unimportant and a linguistically ideal text takes over. Specifically, it becomes important to consider whether the writing and wording of a manuscript accurately expresses the wording and meaning that was intended by its scribe. Perhaps letters have been erroneously omitted or inserted, so that the writing of the scribe does not correctly manifest his intended wording.[42] Or

[41] There is, of course, an element of idealization built into the very notion of transcription itself, inasmuch as the process requires that various markings be identified as tokens of abstract types. At times, this process can involve a kind of reconstruction, especially when a poor scribe has left barely legible handwriting on a manuscript. The more ill-formed a letter is, the more an editor must rely on higher abstractions in order to identify that letter. While this process is akin to graphetic reconstruction, it is better spoken of simply as the deciphering of barely legible handwriting. It has a close parallel in other non-reconstructive interpretive processes such as the expansion of abbreviations and the disambiguation of ambiguous writing and wording.

[42] Orthographic spelling variations serve to helpfully illustrate the difference between reconstructive graphemic transcriptions and rewrites. Although both of these representations might adjust the spelling of an ancient manuscript, they will do so for very different reasons. A reconstructive graphemic transcription will let the spelling of an ancient scribe stand, provided it represents a spelling that would have been deemed acceptable by the scribe himself had he bothered to check over his finished work—but it will correct a specific spelling if it is deemed that the scribe himself would have done so if given the opportunity. So the question relevant to graphemic reconstruction is not ‘what (I think) he should have written’ but ‘what (I think) he would have written if he was writing more slowly and carefully’. By way of contrast, a rewrite will consistently employ spellings in conformity with the conventions prescribed by modern scholarship.

perhaps an overt grammatical error has been made, so that the wording as it stands is nonsensical.[43] In both cases, an editor who makes corrections is concerned with *linguistic errors* and hence a *linguistically* ideal object. He or she is relying upon higher-order linguistic abstractions in order to produce an idealized lower-order linguistic representation.

*Summary and Implications*

Theoretically speaking, the editing of ancient texts is a complex process. Here I am concerned especially with the process of *representation* that is involved in producing an edition. This process is inherently selective, which means that an editor can represent only a sub-set of the total information available to him or her.

The selection of information for representation is complicated by the fact that many different forms of information are possible. Here I have described five default representations. A manuscript may be represented as a *physical artefact* using photographic technologies. Such representations are called *images*. A manuscript may also be represented as *ancient writing*, in which case one must select either a full set of graphs or a more restricted set of graphemes. Such representations are called either *graphetic* or *graphemic transcriptions*. Moving further into linguistic territory, a manuscript can be represented as ancient *wording*. Since this involves re-expressing the wording of an ancient text using a modern form of writing, I have opted to call such representations *rewrites*. Finally, a manuscript may be represented simply as *a meaningful text*, using some modern language as a form of expression. Such representations are called *translations*.

With regard to each of these different representations, it is possible to focus on an *actual* object or an *ideal* object. Graphetic transcriptions, which represent the most concrete of all linguistic objects, must decide whether to take into account the *passage of time*, which frequently damages physical writing. Graphemic transcriptions and rewrites, which represent more abstract linguistic objects, must decide whether to take into account the *scribal error*, which sometimes ‘damages’ a text’s writing or wording.

Given that it is possible to represent a text in such a large number of

[43] Please note that the ideal grammar in view here is that of the scribe himself, not some hypothetically ‘correct’ grammar. Thus the purpose of lexicogrammatical reconstruction is to idealize the actual wording of a text—not to reword in a ‘correct’ form the perceived meaning of a text.

theoretically distinct ways, the practical question must be asked: Is it possible to somehow integrate different kinds of information? For instance, what about producing a representation that includes information about writing and wording and also about physical deterioration and scribal error? The short answer to this question is ‘Yes, sometimes it is possible’. As we will see in a moment, however, a number of considerations need to be taken into account before such an approach is deemed desirable.

## *Making an Edition: Alternative Approaches*

I have just alluded to the most important methodological decision that is involved in the production of a manuscript edition: how to deal with the various levels of abstraction that need to be represented. In this section I will discuss how existing print and digital editions have attempted to represent information from different levels of abstraction. I will show that print editions have integrated multiple levels of abstraction into composite representations, even though this practice necessarily restricts the amount of information that can be represented. I will then show that the EpiDoc Collaborative similarly restricts the amount of information in its digital editions, even though such compromises are no longer necessary. Finally, I will propose an alternative implementation of XML that is designed to overcome the limitations of previous print and digital editions.

### *Traditional Print Editions*

Non-digital editions employ the technology of the printer. For this reason, they are subject to the limitations of the printed page. Images can be reproduced, but they cannot be altered once printed. Typographic content can be arranged in many ways, but it is always fixed in place. Multiple representations can be provided, but they may not fit on a single spread. And, of course, each additional page requires more material and thus increases both the cost and size of an edition. Given these limitations, traditional editions have had to make difficult choices about how best to represent ancient manuscripts.

Generally speaking, print editors have attempted to integrate as much information as possible in as few representations as possible. For this reason, one will rarely find more than three or four of the following representations in a printed edition: an image, a diplomatic transcription, a semi-diplomatic transcription, a reading text and a translation. The first and last of these are uncomplicated. Images can take different forms, and translations can employ different modern languages or translation principles, but in each case there is only a single level of abstraction in view. Moreover, whether or not images and translations appear in an edition is determined entirely by practical, financial or legal factors. More comprehensive editions have them, but not all editions can afford to be comprehensive.

The diplomatic transcriptions, semi-diplomatic transcriptions and reading texts that are found in traditional print editions do not align with the graphetic, graphological and lexicogrammatical levels of abstraction defined in this article. They should not, therefore, be confused with the three 'ideal' modes of representation defined above (i.e. graphetic transcriptions, graphemic transcriptions and rewrites). Generally, diplomatic transcriptions prioritize the faithful representation of a manuscript's writing (i.e. graphetics and graphology) whereas reading texts prioritize the accessible representation of a manuscript's linguistic content (i.e. lexicogrammar). In between the two, semi-diplomatic transcriptions attempt to encompass graphetic, graphological and lexicogrammatical information in a single composite representation. This blurring of theoretical distinctions can be partly explained by the fact that readers do not wish to be constantly flipping back and forth between different pages in order to obtain different kinds of information. But it is also due to the practical constraints that inevitably limit the scope of every print edition. Just as not all editions include images and translations, not all editions include three distinct original language texts. Therefore, in order to understand fully the original language texts published in a given edition one must consult the introduction of that edition to learn the purposes and principles that have guided the creation of its text(s).

What about the Leiden conventions? Over time, the Leiden conventions have emerged as a standard way to indicate editorial reconstructions, expansions and emendations. In essence, they define a subset of graphetic, graphological and lexicogrammatical information, and they clarify how that information should be indicated visually. Of course, the conventions do not dictate the number of texts required by an edition or the content that should be supplied in each text. Editors must still decide *if* and *where* to present the information prescribed by the Leiden conventions. Further, editors must decide *if, where* and *how* to present the information that is *not* encompassed by the conventions. Arguably, however, the Leiden conventions have made it much easier for print editions to supply only a single edited text, with the result that diplomatic transcriptions and reading texts are less common.

So then, in the production of print editions it has never been possible to be exhaustive; rather, it has been necessary to decide which forms of information are most important and how they can be visually displayed in a usable manner. Editors have decided *which* information to transcribe. They have then decided how to present *that* information visually. Because readers do not wish to flip back and forth manually between distinct representations, conventions have been developed that permit the most frequently sought information to be presented in a standard format in a single place. The result is that the most popular representation employed by print editions is a composite text that represents graphetic, graphological and lexicogrammatical information, encompassing both observation and reconstruction. Diplomatic transcriptions and reading texts are used primarily when an editor wishes to overcome specific limitations of the single text approach. Irrespective of whether individual transcribers and editors possess a clear understanding of the information that is *theoretically available* for representation, the form of a traditional print edition is determined by the *practical goals* of each specific transcription project.[44]

### *EpiDoc Digital Editions*

With digital editions, the fixity of the printed page has been overcome, and this monumental revolution has opened up new horizons in the preparation of scholarly editions.[45] The nature of XML is such that

[44] In saying this I wish to deny the suggestion that traditional practices are theoretically unsound. To the contrary, traditional practices often reveal a sophisticated awareness of the textual objects they study and represent. Daniel O'Donnell correctly observes that 'While the limitations of the printed page have undoubtedly dictated the form of many features of the traditional critical edition, centuries of refinement—by trial-and-error as well as outright invention—also have produced conventions that transcend the specific medium for which they were developed. In such cases, digital editors may be able to improve upon these conventions by recognising the (often unexpressed) underlying theory and taking advantage of the superior flexibility and interactivity of the digital medium to improve their representation' ('Back to the Future: What Digital Editors Can Learn from Print Editorial Practice', *Literary and Linguistic Computing* 24 [2009], pp. 113-25 [115]). Similar things can be said about the production of epigraphic and papyrological editions.

[45] In addition to the special issues cited in note 37, I direct interested readers to: Peter Robinson, 'Where We Are with Electronic Scholarly Editions, and Where We Want to Be', *Jahrbuch für Computerphilologie* 5 (2004), pp. 123-43. Online: http://computerphilologie.uni-muenchen.de/jg03/robinson.html; Peter Robinson, 'Current

new information can be added at any time, while collaborative editing environments like the Papyrological Editor facilitate large-scale and long-term projects. This necessitates something of a paradigm shift in that it is no longer helpful for the format of an edition to be determined by the immediate goals of a single editor or project. To the extent that online editing projects are *open* and *ongoing*, scholars must decide to think very broadly about what information might be represented in an edition and how all of that information might be effectively encoded.[46] The EpiDoc standards constitute a good step in this direction.

In my earlier discussion of EpiDoc, I observed that the scheme employs XML because the use of this more abstract markup permits a separation of structure and presentation. In other words, the fixity of the printed page is overcome because the information that is structured within a static XML document can be rendered visually in countless different ways. Where a diplomatic transcription is desired, this can be achieved through an XSL stylesheet. Where a semi-diplomatic transcription is desired, this can be achieved through the use of a different stylesheet. At any point new stylesheets can be designed, each selecting a new subset of the total information encoded and each selecting a way of displaying that information. This is the digital equivalent to page turning, except

Issues in Making Digital Editions of Medieval Texts—Or, Do Electronic Scholarly Editions Have a Future?', *Digital Medievalist* 1 (2005), n.p. Online: http://www.digitalmedievalist.org/article.cfm?RecID=6; Gabriel Bodard and Juan Garcés, 'Open Source Critical Editions: A Rationale', in Marilyn Deegan and Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World* (Farnham, UK: Ashgate, 2009), pp. 83-98; Greg Crane, 'Give Us Editors! Re-Inventing the Edition and Re-Thinking the Humanities', in McGann, Stauffer and Wheeles (eds.), *The Shape of Things to Come*, pp. 81-98.

[46] Thus the important question is not what can be encoded immediately, but what might be encoded eventually. As D.C. Parker somewhat humorously observes, 'It is almost inevitable that the new electronic transcriber will soon become taken with a desire to represent everything that is visible on the page—ink marks that might be smudges, stained areas, possible spaces in the text, letters of an unusual size or shape or out of alignment, changes in ink colour, ligatures. This desire will in time give way to a recognition that this aim cannot be achieved and be replaced with a pragmatic recognition that the main virtues are consistency and accuracy in representation of the most important data. It will always be possible for someone to add more detail at a later date' (*An Introduction to the New Testament Manuscripts and their Texts* [Cambridge: Cambridge University Press, 2008], pp. 104-105). Of course, whether it will be easy to add more detail at a later date depends on how an edition has been designed in the first place.

that nobody needs to produce each of the different pages manually. With XML, *every* encoded manuscript can be *automatically* transformed and displayed using a newly produced stylesheet. And as the InsAph project attests, moving between these displays can be as easy as clicking a button.

But what can be said about EpiDoc's particular implementation of XML, especially as regards the theoretical distinctions outlined in this article? Do the Epidoc standards provide a means by which a group of collaborating scholars can encode in detail all of the important levels of abstraction and domains of reconstruction? And are the various levels and domains integrated in a way that facilitates computational analysis and the creation of flexible interfaces? The short answer in both cases, unfortunately, is 'No'. Here is the main reason: the EpiDoc scheme elevates the identification of graphemes and the recording of page layout to a position of high importance but disregards most other aspects of graphology and graphetics. Does this sound at all familiar? It should. After all, this is the necessary compromise that was accepted by epigraphers and papyrologists for the production of affordable and manageable print editions.

Demonstrating the pervasiveness of this compromise within EpiDoc is somewhat more complicated than pointing it out in a printed text, but it is not overly difficult. Anne Mahoney has said about EpiDoc that 'The basic philosophy of the guidelines…is clear. The simplest rule is that whatever is actually on the stone is in the content of the elements, while editorial changes and additions are in attributes.'[47] One imagines a carefully encoded base of graphetic content, which is then overlaid with markup encompassing all of the other aspects of an edition, including physical deterioration and visibility, graphemic distinctions, abbreviations, scribal errors and editorial emendations, etc. And one imagines that the rendering of this information in a modern, readable format is accomplished somehow through the use of a stylesheet. In practice, however, palaeography gets a very short shrift, and the Leiden approach is adopted. 'Spacing, punctuation, and capitalization are all added or adjusted silently, as are accents and breathings for Greek. Readers are generally simply expected to know that ancient writing conventions are different from modern.'[48] In other words, the EpiDoc community has built modern writing conventions

[47] Anne Mahoney, 'Epigraphy', in Lou Burnard, Katherine O'Brien O'Keeffe and John Unsworth (eds.), *Electronic Textual Editing* (New York: Modern Language Association of America, 2006), pp. 224-40 (234). Online: http://www.tei-c.org/About/Archive_new/ETE/Preview/mahoney.xml.

[48] Mahoney, 'Epigraphy', 226.

into the foundational content of its editions in such a way that there is no easy way for graphetic and graphological information to be added later without compromising the basic philosophy of the guidelines. It has taken the information recorded in traditional print editions and encoded that information in XML, but it has not created a document structure designed to integrate other information.[49] Graphetic and graphological information might be encoded in a separate document division, but such a solution merely perpetuates the fragmentation that characterizes print editions.

Why has EpiDoc taken the approach that it has? I suspect the main reason is EpiDoc's conscious concern for backwards compatibility, an admirable stance that has permitted the creation of such valuable things as the Chapel Hill Electronic Text Converter. But while it is natural that a new digital encoding scheme should seek to import existing data, it does not follow that digital editions should limit themselves to the data recorded in traditional print editions. Recall the wisdom of Busa: 'It would be much better to build up results one centimetre at a time on a base one kilometre wide, than to build up a kilometre of research on a one-centimetre base'. Recall also Busa's insistence that digital technology should not merely 'enhance' existing research but must open up new avenues of exploration that can serve to 'extend' the scope of humanities research.[50] In the case of traditional print editions, it was never feasible to perform large scale analyses of graphetic and graphological information, so the

[49] Such a perspective on the function of EpiDoc is evident in some of the comments of its developers. For example: 'The XML version therefore should not be viewed as a replacement for Leiden, which is easier for scholars to produce and to read, but as an interchange format to be used when Leiden needs to be read, manipulated, or transmitted by a computer' (Hugh Cayless *et al.*, 'Epigraphy in 2017', *Digital Humanities Quarterly* 3.1 [2009], §19). Online: http://www.digitalhumanities.org/dhq/vol/3/1/000030/000030.html.

[50] Writing in 2003, Peter Robinson made the following comment: 'Let us observe two things missing from almost all electronic scholarly editions made to this point. The first missing aspect is that up to now, almost without exception, no scholarly electronic edition has presented material which could not have been presented in book form, nor indeed presented this material in a manner significantly different from that which could have been managed in print… The second missing aspect of most electronic scholarly editions relates to their failure to use new computer methodologies to explore the texts which they present… The only tool many editions add is text searching—and many do not even provide that. Very often too computerized tools are not used in the preparation of the editions: a database might sometimes be used for gathering some data, but that is all' ('Where We Are with Electronic Scholarly Editions', §6).

absence of this information was largely inconsequential. But scholars are now beginning to perform large scale quantitative analyses using digital editions, and the progress of such efforts depends on the creation of more comprehensive data sources than are presently available. In the case of the older editions now available in EpiDoc XML, new information might be added using collaborative, online editing environments. The EpiDoc standard, however, does not provide a way for this information to be effectively encoded in its existing XML structure. Nor will Leiden+ provide a useful mechanism for inputting this information. Moreover, it is not self-evident that the standard will be easily adaptable to the needs of OCR technologies in the case of inscriptions and manuscripts that are being edited for the first time.

To summarize: while the data contained in traditional editions can be helpfully converted to XML using the EpiDoc standard, this information represents only a small portion of the information that is relevant to the study of ancient inscriptions and manuscripts. This being so, it must be considered whether the EpiDoc standard can become comprehensive enough to encode all of the data that might be of interest to future scholars. I do not claim to have a definite answer to this question. But to the extent that EpiDoc has been designed with reference to the traditional distinctions made in print editions rather than the theoretical distinctions that define the data space available for exploration, I favour the development of a more robust encoding scheme that is designed to handle all of the information that might someday be desired in a comprehensive digital edition. Such a scheme should be able to import the information that is already available in print editions or EpiDoc editions, but it should restructure that information in such a way that additional detail can be added, not only with respect to the linguistic levels of abstraction defined in this article, but also with respect to any additional perspectives that might arise in the future.

### *Modular Digital Editions*

Much like the edited texts that are found in print editions, the EpiDoc scheme attempts to encode several different strands of information in its ‘edition’ division. Conversely, it places images, diplomatic transcriptions and translations in separate document divisions. Here I will very briefly sketch an approach to digitization that expands the number of document divisions to at least five. This approach defines a distinct document division for each level of abstraction, and it carefully designs these divisions in order to account for two kinds of information: (1) information that is characteristic of each specific level; and (2) information that is relevant to the interpretive processes involved in moving up to that level. In addition, I will propose that Xlinks should be used to integrate distinct document divisions in a manner reminiscent of stand-off annotation. Because of these links, a modular approach not only allows a broader range of information to be encoded at a higher degree of specificity, it also allows multiple levels of abstraction to be integrated by searches or stylesheets. Ultimately, it permits an almost infinite variety of visual representations, and it makes it possible for users to easily navigate between those different representations.

Before describing the creation of an edition along these lines, I wish to reiterate the fact that an XML document is itself a mode of representation, albeit one that looks like nonsense to uninitiated observers. In contrast with traditional print representations, which are fundamentally visual, an XML edition is a kind of database, a repository of information, that can be accessed in various ways. It is a digital storehouse that can be consulted by a search engine, or analyzed computationally, or drawn upon as a resource for the creation of visual modes of representation. If more information is encoded in an XML document, more information is available for searches, analyses and displays. The trick is to manage information well so that it can be efficiently encoded and accessed. While a well-organized XML document may not look at all like an image, or a transcription, or a rewrite, or a translation, it will have the potential to be *transformed* into these other modes of representation.

As discussed earlier in this article, two theoretical dimensions define the information that must be organized in a digital edition. First, there are various levels of abstraction. There is the physical appearance of a manuscript, and there are the various objects defined by graphetics, graphology, lexicogrammar and semantics. Secondly, alongside the observable manuscript itself, there are various editorial idealizations that must be accounted for: reconstructed graphs, corrected writing and corrected wording. My proposal is that these two dimensions should be handled in two complementary ways. Levels of abstraction should be encoded in separate document divisions, with Xlinks used to integrate those divisions in a manner akin to stand-off annotation. Conversely, graphetic reconstructions, graphological emendations and lexicogrammatical emendations should be encoded within their respective document divisions using inline elements and attributes. It is well beyond the scope of this article to present the actual XML underlying my proposal. (And in any case, it is unlikely that XML samples would be of great interest to most biblical scholars, epigraphers and papyrologists.) Instead, therefore, I will briefly explain how my proposal would affect the production of a digital edition.[51] The technical details, as essential as they are, must await another publication.

In simple terms, the production of a modular digital edition resembles the production of a fully comprehensive print edition in the sense that multiple representations are created and marked up, with more abstract information emerging from the examination of less abstract information. A hypothetical scenario might unfold something like this: A new manuscript is digitally photographed, perhaps in various different ways, and the resulting images are documented in the 'image' division of an XML document. Those images are then scanned using Optical Character Recognition software. The OCR software identifies discrete markings on the manuscript, encodes those markings as discrete elements in a 'graphetic' division, links each individual graph to the relevant spatial zone of the scanned image, assigns to each graph element any palaeographical attributes that have been identified in the scanning process and attempts to structure the graphs according to their page layout.[52] As a result of this initial process, an editor possesses a graphetic transcription containing information about the specific letter forms visible on a manuscript. Moreover, this transcription is linked to an image in such a way that the editor needs only to click on a specific graph in order to call up a photograph of it. Alternatively, an interlinear display might be automatically generated, with the content of the 'graphetic' division

[51] For a much more detailed discussion of the editorial process and how computer technology might assist that process, see Melissa M. Terras and Paul Robertson, *Image to Interpretation: An Intelligent System to Aid Historians in Reading the Vindolanda Texts* (Oxford: Oxford University Press, 2006); Ségolène M. Tarte, 'Papyrological Investigations: Transferring Perception and Interpretation into the Digital World', *Literary and Linguistic Computing* 26 (2011), pp. 233-47.

[52] In order to gain some sense of how digital editions might integrate with the work of palaeographers, interested readers should consult Melissa M. Terras and Paul Robertson, 'Downs and Acrosses: Textual Markup on a Stroke Level', *Literary and Linguistic Computing* 19 (2004), pp. 397-414; Murray McGillivray, 'Statistical Analysis of Digital Paleographic Data: What Can It Tell Us?', *Digital Studies/Le champ numérique* 0.11 (2005), n.p. Online: http://journals.sfu.ca/chwp/index.php/chwp/article/view/A.33/54. Ciula, 'The Palaeographical Method'; Peter Stokes, 'Computer-Aided Palaeography, Present and Future', in Rehbein, Sahle and Schaßan (eds.), *Kodikologie und Paläographie im digitalen Zeitalter*, pp. 309-38. Online: http://kups.ub.uni-koeln.de/volltexte/2009/2978.

positioned above strips of the underlying image. Using these resources, editors carefully check and correct the graphetic transcription, perhaps by selecting from alternative graphs proposed by the OCR software itself. They also ensure that the transcription has been correctly structured using page layout elements. They then add additional information about the physical status of each encoded graph, perhaps by selecting one or more graphs and then clicking through a system of pre-defined attributes. Finally, editors may manually encode graph elements for illegible, deleted or damaged graphs, in each case marking up the newly created elements with information about their physical status or about the reconstructive process itself. The 'graphetic' division obtained at the end of this process will not visually resemble a graphetic transcription, but it will contain all of the information needed in order to generate a graphetic transcription. As a bonus, each element in the 'graphetic' division links back to an underlying zone in the 'image' division.

Along similar lines, the encoding of a 'graphological' division involves the scanning of a 'graphetic' division. Each uniquely identifiable graph element in a 'graphetic' division is converted into one or more uniquely identifiable grapheme elements, and links are created between the new grapheme elements and the underlying graph elements. Moreover, the newly created grapheme elements are automatically marked up with information about the specific modifications or expansions carried out. Thus the resulting 'graphological' division explicitly identifies re-expressed ligatures and expanded abbreviations, making it easy to locate or analyze these phenomena, easy to render them in the manner of a traditional semi-diplomatic transcription, and even easy to call up the underlying graph elements. After checking this information carefully, editors encode information pertaining to scribal corrections and scribal error. Marginal and interlinear corrections, which were encoded simply as marginal and interlinear markings in the 'graphetic' division, are now encoded as alternative readings and marked up with informative attributes. In a similar manner, blatant but uncorrected graphological errors might be corrected and classified. Just as with graphetic reconstructions, the inclusion of scribal corrections and editorial reconstructions does not at all damage the integrity of the edition because the inline elements and attributes used to encode this information add new content without replacing existing content. Scribal corrections and editorial reconstructions are thus available to searches, computational analyses and stylesheets but are not imposed upon them.

Moving from a 'graphological' division to a 'lexicogrammatical' division involves the identification of individual words, the introduction of modern graphological conventions and the insertion of chapter and verse references. An automated scanning process might be used to create word elements in a new ‘lexicogrammatical’ division, which would then link down to underlying graphemes in the ‘graphological’ division. Alternatively, editors might import the basic characters in a ‘graphological’ division and then manually encode word divisions. Either way, after a segmented text has been prepared, editors regularize the content of each word element. They supply diacritics, accents, punctuation, mixed case, etc., and they standardize the spelling of the text. They then structure the text into sections, paragraphs or verses, and (if required) insert chapter and/or verse references. They encode editorial emendations to the wording of the text using inline elements, once again leaving the unemended text untouched in the process. Finally, they might encode such additional details as onomastic or lexicographic information, perhaps by means of an automated parser that tags word elements with the necessary attributes.[53] Because the resulting ‘lexicogrammatical’ division links down to underlying graphological, graphetic and even visual data, the edition as a whole contains all of the information required to produce a fully regularized reading text, with the added bonus that it is possible to visually indicate each editorial intervention so that it can be peeled back in order to reveal the underlying evidence relied upon by the editors.

The creation of ‘translation’ divisions is very straightforward. Using an interlinear display, editors translate the lexicogrammatical content of a text in such a way that the resulting translation links back to specific word elements in the underlying ‘lexicogrammatical’ division. Although many scholars will begin their exploration of an ancient text using a modern translation, by means of the linking mechanism proposed here it will be very easy to call up the original language text. While looking at a text, a scholar might decide to view information about editorial reconstructions and emendations (perhaps even using the Leiden conventions). He or she might then drill down into lower levels in order to see the concrete evidence that underlies specific editorial decisions. In essence, end-users will not only have distinct representations, each carefully defined and managed; they will also have the option to display visual signals that will inform them about the steps of abstraction involved in the encoding

[53] For a discussion of this process from an EpiDoc perspective, see Gabriel Bodard, ‘Digital Epigraphy and Lexicographical and Onomastic Markup’, n.p. Online: http://www.stoa.org/archives/1226.

process, thereby alerting them to potentially interesting data available in an underlying level. Nobody can visually process all of the information stored in a comprehensive XML edition at one time, so the important thing is to carefully integrate different strands of information. This is the vision that has guided me towards the modular approach described here.

As a final note, I wish to point out that the complexity introduced by the need to employ uniquely identifiable elements is offset by the resulting simplicity of the encoding process. This is a vitally important consideration, given that few classicists or epigraphers or papyrologists or biblical scholars will learn how to manually encode in XML.[54] The Papyrological Editor addresses this obstacle by means of its Leiden+ encoding language, which visually resembles what scholars are used to seeing (and producing) in print editions. However, an even more elegant solution is possible by means of a modular approach, since editors will be able to edit or mark up the content of each division using a simple graphical user interface. This is because virtually all editorial actions entail one of the following simple adjustments: assigning an attribute to a uniquely identifiable element, changing the content of a uniquely identifiable element or inserting a new uniquely identifiable element. All of these adjustments can be performed directly on a rendered text using fields and check boxes. As regards the encoding of *alternative* content such as variant readings or editorial emendations, a simple wizard can be designed that inserts new content while simultaneously assigning attributes to each alternative. A display toggle can then switch back and forth between any available alternatives so that they can be edited directly as before. Granted, expertise will be required on the part of any scholar who wishes to design an engine that will automatically generate parsed levels. Yet the manual creation of a new level can be achieved through a simple import option that copies the content of one level into a blank template and creates the necessary links. Moreover, simple changes to the links in a modular edition can be made in a manner that should prove recognizable to anyone who has ever inserted a hyperlink into a word processing document. A modular digital edition, therefore, provides extended breadth and greater flexibility, while making it easier for non-technical scholars to participate in collaborative editing projects.

[54] Robinson, having discussed the continuing dominance of print editions, states the following (with which I heartily agree): ‘Our goal must be to ensure that any scholar able to make an edition in one medium should be able to make an edition in the other’ (‘Current Issues in Making Digital Editions of Medieval Texts’, §25).