5

EpiDoc: Epigraphic Documents in XML for Publication and Interchange

Gabriel Bodard

EpiDoc History

Epigraphy, the study of texts inscribed or incised on durable materials, is a subdiscipline of Classics with a long history. We know that Byzantine scholars copied down ancient texts they discovered on their travels (much of the *Palatine Anthology* is made up of such texts, for example). As Classics matured as a discipline in the nineteenth and early twentieth centuries, recognized conventions developed to indicate the condition of texts of inscriptions, just as they did in the related field of manuscript transcription and, at the end of the nineteenth century, in the new field of papyrology. For example, square brackets around a sequence of characters usually indicated that these letters were missing from the stone (or papyrus) and had been restored by the modern editor, either by comparison with existing texts or by some other speculative means.

These conventions were not standardized, but they were usually clear enough. At any rate it was obvious to most readers which characters were on the ancient material and which were supplied, although it was not always entirely transparent why they were absent in the first place (lost due to damage, lost due to erasure, omitted in error, or omitted for abbreviation, for example). Since most nineteenth-century epigraphic publications would have included a drawing, facsimile, or diplomatic transcript of the text, it was in any case relatively easy to deduce the state of the original text and the editorial decisions involved in creating the interpretive copy.

Due to changes in both publishing and scholarship itself in the early twentieth century, these *ad hoc* conventions—which were adjusted and elaborated by many editors who found them almost but not quite suitable to their needs—came to be seen as inadequate for the publication of ancient writing. Among other things, the creation of large corpora of inscriptions, such as the *Corpus Inscriptionum Latinarum*, made desirable, even essential, the use of consistent, universal, agreed standards for epigraphic conventions. The Leiden Convention, a meeting of international scholars in 1931, aimed to draw up just such a universal standard, and to a large extent it succeeded.[1]

In the Leiden style of epigraphic transcription, a pair of square brackets ('[' and ']') always signifies that text has been lost from the stone or papyrus due to some kind of physical damage. Letters inside the square brackets have been restored by the editor; dots or dashes inside the square brackets denote letters that cannot be restored at all, and may be of unknown extent. Similarly coherent conventions exist for the use of parentheses (expansion of abbreviation), angle brackets (omitted letters), curly braces (superfluous letters), subscript points (unclear or ambiguous letters), and so forth. With the exception of those texts published before 1931 (and a few editors who rejected or modified Leiden[2]), it was now possible for almost any epigraphic or papyrological edition to be read by a scholar familiar with Leiden, leaving no doubt as to which letters were restored or corrected and for what reasons.

There are, however, a few variations in the use of Leiden between papyrologists, on the one hand, and epigraphers on the other; several scholars have proposed updates or modifications to the conventions.[3] There are, for example, the upper-half-square-brackets used by many papyrologists and some epigraphers (principally but not exclusively Latinists) to indicate letters erroneously substituted or incorrectly executed, and corrected by the editor; in original Leiden this condition was indicated more ambiguously with angle brackets which could also indicate the restoration of omitted letters. Most epigraphic and papyrological volumes still contain a page of "editorial conventions" for clarity.[4] Further explanation and modification of the conventions became necessary as ancient texts of this kind were stored in computer databases and other electronic formats such as XML (Extensible Markup Language, a global standard, on which see further below and n. 11). In the early days of large, publicly accessible databases of Greek and Latin there were two main strategies for encoding texts and *sigla*. The options were either to employ Leiden, using the common brackets but not the more rare symbols, subscript points, or underlining, or to produce a more complex system that used combinations of brackets and other ASCII symbols to represent all of the typographic features normally used in the encoding of such texts.

The former strategy was often adopted by the database projects, especially in the days before the wide acceptance and compatibility of Unicode. This led to a slight devaluation of Leiden, perhaps, but nevertheless the texts were easy enough to read and search, and scholars could in any case refer to the original publications in case of uncertainty. The latter strategy is exemplified by the case of Beta Code, a system devised in the early 1980s and used by projects such as the TLG and PHI digital libraries. Beta Code is both an encoding scheme for non-Latin characters (principally the Greek alphabet) and a markup scheme to represent the various typographical features of a text: brackets and other *sigla*, lineation, pagination, fonts, and so forth, all using combinations of ASCII characters. In its most sophisticated incarnations, Beta Code is an extremely powerful and comprehensive scheme for the encoding of Greek and Latin texts.[5] The only real weakness of Beta Code as an encoding scheme is its idiosyncrasy; it is not a widely recognized or supported encoding. In order to display Beta Code as Greek (or Latin) text formatted as intended—much less to process and search it intelligently—one requires a highly specialized piece of software that only exists in a few places. There are widely accepted schemes for encoding and marking up this sort of information, namely Unicode and XML, which are not unique to classicists or even academics. Accordingly, many tools exist to process and display these technologies; they are supported as standard in all operating systems and almost all software packages. A lot of very wealthy industries have vested interests in making sure that transfer and upgrade to future technologies are as smooth as possible. It obviously makes sense for epigraphers (and classicists in general) to piggyback on such technologies rather than trying to reinvent the wheel with our much more limited resources.

In 1999 a commission on Epigraphy and Information Technology held a round-table meeting, convened by Silvio Panciera, under the auspices of the Association Internationale d'Epigraphie Grecque et Latine. Among the outcomes of this meeting, in addition to an alliance between the major Latin databases at Heidelberg, Rome, and Bari,[6] was the statement that the data in these databases needed to be (a) in Unicode, and (b) archived in XML.[7] In response to this report, Tom Elliott, then director of the Ancient World Mapping Center at the University of North Carolina at Chapel Hill, made public the EpiDoc Guidelines, recommendations for XML mark-up of epigraphic documents, that he and colleagues had been working on privately for some time.[8]

These guidelines and other tools have since matured considerably through extensive discussion in online fora,[9] at several conferences, and through the experience of various pilot projects. The first—but not by any means the only—major epigraphic project to adopt and pilot the EpiDoc recommendations was *Inscriptions of Aphrodisias*.[10] In the course of this process the guidelines and tools have reached a degree of maturity and stability for the first time.

EpiDoc Philosophy

EpiDoc specifies the use of XML, Extensible Markup Language, an industry standard maintained and documented by the World Wide Web Consortium for communication and storage of structured data.[11] XML is a software- and platform-independent language, optimized for compatibility, interchange, and durability, which means that it is ideal for archive storage as well as web and database publication. Since XML, and its parent language SGML, are used almost universally for encoding and storing data in the commercial sector, by computer professionals, publishers, analysts, archivists, economists, and so forth, advantages over a proprietary database system are undeniably manifest. In particular, it is likely that any changes in technology that require upgrades to either the encoding of XML itself, or its transformation and delivery, will be handled by those with the resources to do so, and that academic projects can coat-tail on this progress, rather than having to invest in expensive solutions themselves or see their materials fall out of date.

XML, unlike many mark-up and publishing systems (including HTML on the Web and RTF—Rich Text Format used by word processors) does not merely encode the appearance of a text, but can also embed information about its structure and semantics. Appearance in any given form, whether a web page, a printed text, or an audio version for the blind, will be handled by a set of stylesheets (a computer file that defines how to convert an XML document into some other digital format). The stylesheet can be instructed, for example, to separate paragraphs by a blank line, to render foreign words in italic face, or to put square brackets around editorial supplements. This technology can also sort elements in a given order, treat them differently based on context, index certain types of keywords (such as those foreign words, but not, say, titles or other words in italics), create tables of contents based on date, genre, or some other category, pull the data into a larger corpus of similar materials, and many other transformations.

Because XML allows for structured and semantic markup, it can not only be used to encode data for display or publication, but can be processed, queried by a search engine, or translated into another markup or database system.

XML is almost infinitely customizable, with each instantiation being defined in a schema file (either DTD, Document Type Definition, or latterly a RelaxNG or W3C Schema), which provides a menu of tags and attributes, and specifies the contexts in which they may occur. Rather than completely reinvent the wheel, and so as to be compatible with established standards, EpiDoc is built using a subset of the XML defined by the Text Encoding Initiative (TEI). This schema, a widely used XML system in the fields of literature and linguistics, is particularly suited to the transcription and description of texts and manuscripts.[12] Using a TEI schema maximizes the compatibility of EpiDoc encoded inscriptions with other text projects in the humanities generally. The EpiDoc Guidelines, therefore, rather than being an entirely new system, may be considered as a local guide to practice within the larger TEI guidelines.

An essential concept behind EpiDoc is the understanding that this form of semantic markup is not meant to replace traditional epigraphic transcription based on the Leiden conventions. XML may (and almost inevitably will) encode more information than the range of brackets and sigla used in Leiden, but there will always be a one-to-one equivalence between Leiden codes and markup features in the EpiDoc guidelines. This means that a text encoded in Leiden can always be marked up using EpiDoc XML with very little extra editorial intervention—in fact tools exist whereby this process can be almost entirely automated.

An EpiDoc file is a representation in XML of the edition of one inscription or a group of inscriptions. The minimal file will contain a text in Greek or Latin, probably with editorial sigla. It may also contain apparatus, translation, commentary, description, and dating of the text or object, history of the inscription, bibliography, or any other information that is normally published in a scholarly edition. The file may also contain cross-references to other texts, files, indices, tables, appendices, and images. Since XML is more flexible than a database structure, it may also contain, wherever text occurs, any number of tagged terms, keywords, or names for indexing; indices and tables of contents may then be generated by stylesheets as one set of the outputs from the XML. Any file that contains the bare minimum of Greek or Latin text, with all of the distinctions traditionally indicated by the Leiden conventions marked unambiguously in EpiDoc XML, may be considered Leiden-conformant EpiDoc. This XML, since it only needs to contain the epigraphic text, may appear as a fragment of EpiDoc XML within a different schema or database field, so long as it conforms in isolation to the EpiDoc schema.

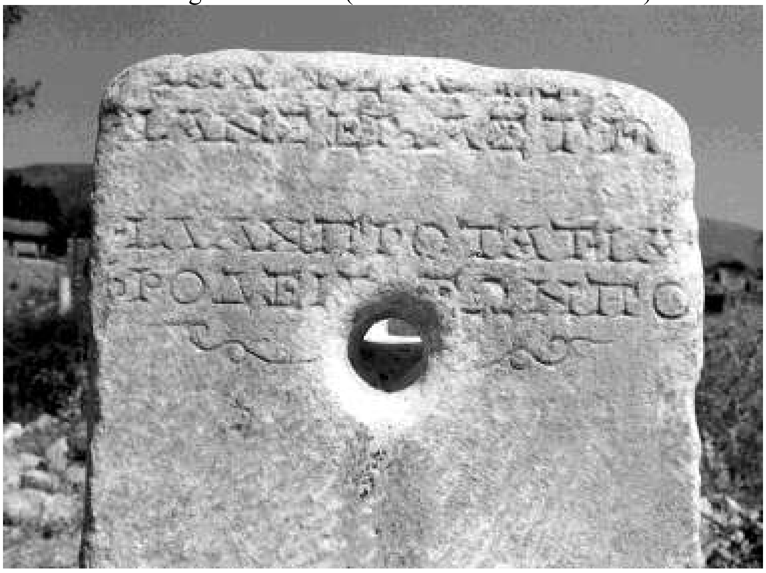
A second level of EpiDoc conformance has also been defined. An XML file that is valid according to the EpiDoc schema and contains both Leiden-conformant transcribed text and enough supplementary information tagged in accordance with certain rules to allow automatic conversion to the database formats of the Heidelberg, Roma, or Bari databases (as appropriate), may be considered EAGLE-Conformant EpiDoc. The recognition of this level of conformance is essential to EpiDoc's role in the epigraphic community as an interchange medium. By providing protocols as well as technologies by which epigraphic data may be moved from various canonical sources into the neutral medium of XML, and correspondingly out into the major community databases, we are not only fostering a degree of *amicitia* but defining the route by which archival versions can be made of all of these texts and collections, so that they remain both available and useful to scholarship.

It is also central to the EpiDoc philosophy that the community provides concrete assistance as well as recommendations. Projects utilising EpiDoc for the publication of inscriptions or papyri are encouraged (although not required) to share with the community their experiences, any enhancements or modifications to the schema or guidelines, and code written or tools created. All tools, code, and stylesheets produced by the EpiDoc Collaborative are made available via the SourceForge repository under the GNU General Public Licence.[13] In addition to the Guidelines, the following tools are explicitly offered as part of the EpiDoc project site: a Web-Application; sample EpiDoc XSL stylesheets; the Chapel Hill Electronic Text-converter (CHET-C); and the Crosswalker tool. These tools are discussed in more detail below, but I shall give a brief description of them here.

The Web-Application is a set of XML files, XSL stylesheets, and site-map files that may be downloaded and run within a free web publication framework called Cocoon.[14] This application contains basic EpiDoc information, and the minimum structure necessary to run web-based elements such as the EpiDoc Guidelines, sample stylesheets, and CHET-C, in a dynamic environment. All of these elements may then be modified and expanded at will, creating a minimal but serviceable development environment. Within this, the modular and adaptable standard EpiDoc stylesheets may be installed, providing the capability to simply convert any set of EpiDoc XML files (those belonging to a project, or the example files provided by EpiDoc) into publishable HTML or other formats.[15] The Chapel Hill Electronic Text-converter is a JavaScript tool that allows a section of Leiden-transcribed text to be pasted into a web form and returns a valid XML marked-up version of the transcription.[16] The Crosswalker tool is a more sophisticated conversion mechanism that can be customized to convert between any structured schema (such as XML or a database) and EpiDoc compliant XML, in both directions. Customization files have been designed and documented so as to be as easy to learn and use as possible.[17]

EpiDoc Examples

Figure 1: *ALA* 2 (© 1980 Mossman Roueché)



As mentioned above, in its simplest form the epigraphic text marked-up in EpiDoc XML contains no more or less detail and complexity than the traditional Leiden transcription. The XML is more verbose, and less attractive to the human eye, but it is designed for a computer to read and process, and the human user should rarely have to write or read the XML without some intervening process or stylesheet. Nevertheless, the principles behind EpiDoc XML are no different from those behind the Leiden *sigla*, and they can be understood with only a little training.

As an example I shall show how one inscription (taken from *Inscriptions of Aphrodisias*) would be marked up in Leiden-conformant EpiDoc XML. The inscription itself can be seen in the photograph (fig 1), followed by the traditional Leiden transcription of the text.

Transcription of *ALA* 2[18]

[Ἰουλίαν Κορνη]-
λ̣ίαν Σαλ̣ων̣[εῖ]-
ναν Σεβαστὴ̣ν
vacat
ἡ λαμπροτάτη Ἀ-

φροδεισ̣[ι]έων πό-
scroll [λις] *scroll*

There are five surviving lines of text on this stone (the last with only scroll-marks surviving), plus one line left blank on the stone (the *vacat*) and a line presumed lost from the top of the text. Characters within the square brackets are entirely lost and restored by the editor; characters with a subscript point beneath them are damaged so that they would be ambiguous outside of their context, but can be read with some confidence by the editor. The first thing to notice about the XML is that all of these *sigla* are omitted: all such semantic information is conveyed using XML tags instead. The first line therefore reads:

<supplied reason="lost">**’Ιουλίαν Κορνη**</supplied>

The characters within angular brackets are the XML "tags", each opening tag having a name (in this case "supplied") and possibly some attributes ("reason", whose value is "lost"); the closing tag, which denotes where the text to which this applies ends, contains the name of the element preceded by a forward slash. All of the characters between these tags are restored by the editor because they were lost from the stone, but any characters immediately following the closing tag are not lost. (In an XML editor—a piece of software like a word-processor that eases the task of editing XML—these elements, attributes, and values may be given different styles or colors to aid the user in seeing what is what. I have simply used bold text for the Greek and normal weight for the tags; XML has no inherent styles but is plain Unicode text, so styling for appearance does no harm.)

Each line of text in the EpiDoc XML, rather than being separated by carriage returns as in a word-processor, is preceded by a tag <lb/> (the trailing slash indicates that this is an "empty" element, marking a place rather than containing text, and does not need to be closed later in the file). The hyphen at the end of the first line is part of the epigraphic conventions for showing that the line break divides a word; this too is represented by markup in the XML (an example of EpiDoc usage that has influenced the TEI recommendation in 2008[19]). The first two lines of this text, therefore, would be tagged:

<lb/><supplied reason="lost">**’Ιουλίαν Κορνη**</supplied>
<lb type="worddiv"/><unclear>**λί**</unclear>**αν**
Σα<unclear>**λ**</unclear>**ω**<unclear>**ν**</unclear><supplied
reason="lost">**εῖ**</supplied>

We can already see that the XML is far more verbose than the Leiden, and far harder to read. Even with the bolding in the text above, it is difficult to make out that the second word of line 2 is the beginning of the name Σαλωνεῖναν. The simple square bracket has been replaced by an XML tag that is 24 characters long, and the subscript point—which took up no horizontal space in the Leiden version—has been replaced by a tag labeled "unclear" before and after each character so marked. Nonetheless, it should be clear that no information has been lost from or added to the text.

Line 3 is almost entirely plain text, although it begins in mid-word and has one damaged letter at the end:

<lb type="worddiv"/>**ναν Σεβαστὴ**<unclear>**ν**</unclear>

Below this, the fourth line of the edition represents a space uninscribed on the stone approximately equal to the height of one line:

<lb/><space extent="1" unit="line"/>

The next two lines are clear enough:

<lb/>**ἥ λαμπροτάτη Ἀ**

<lb type="worddiv"/>**φροδει**<unclear>**σ**<unclear><supplied reason="lost">**ι**</supplied>**έων πό**

The final line, which on the monument has been pierced through due to modern reuse of this stone as a well-head, contains two non-text glyphs to either side of where the restored text obviously once stood.

<lb type="worddiv"/> <g type="scroll"/> <supplied reason="lost">**λις**</supplied> <g type="scroll"/>

All of the text marked up above represents the minimum, exact and reversible conversion of the Leiden *sigla* into EpiDoc XML tags. This conversion can be reliably performed in both directions by existing software—for example, CHET-C for the Leiden to EpiDoc conversion, standard XSL Transformations for EpiDoc to printable Leiden (on both of which see below)—but could also be performed by a human editor with minimal training. The text above, if collected into a single file and surrounded by <ab> (arbitrary block) and </ab> tags, would be valid, Leiden-compliant EpiDoc XML.

Some projects might also choose to add lexical, onomastic, prosopographical, and other information to the XML, either programmatically or by hand. The XML principles at work here are exactly the same as those demonstrated above, although in this case additional information is being inserted into the text, rather than only reflecting the distinctions drawn by Leiden. These tags are more likely to be used for indexing or searching than to affect output rendering, although it would of course be possible to choose to format names differently from other words, for instance. I shall give only a couple of brief examples here.

<w lemma="λαμπρός">**λαμπροτάτη**</w>

In this case the word on the inscription, the "token" in linguistic terms, has been explicitly delimited by the <w> and </w> tags, and also been given a lemma attribute containing the dictionary headword (or "type") to be used for look-up.

<name ref="#Σαλωνεῖνα">**Σαλωνεῖ**<lb type="worddiv"/>**ναν**</name>

Here again a word, in this case a proper name, has been delimited and identified. The regularized form of the name (usually the nominative) is stored in an onomastic database or “authority list” and pointed to by a key on the "ref" attribute, and may be used for indexing or searching, or to link to an external resource for Greek names. Note that (a) only the single name "Salonina" is tagged—the full "Julia Claudia Salonina Augusta" may also be tagged as a person reference and linked to prosopographical information or separately indexed; and (b) the presence of XML tags within the marked-up word or name is not an obstacle to indexing or collating the text itself.

Transformations

The value of XML lies in its capacity to be transformed into a variety of outputs for machine- or human-processing. Most such transformations are carried out by means of scripts written in the Extensible Stylesheet Language (XSL), itself a flavor of XML. These XSL Transformations (or XSLT) may be used to turn the semantic encoding of an EpiDoc file into a web page in HTML, into XML that recognizes a different schema, or into a tabular form for import into a database or other software. Or it may be aggregated with other files into an index, table of contents, concordance, authority list, or other summary. An inscription or corpus of inscriptions marked up in EpiDoc XML may therefore be rendered for display or publication in a variety of forms; indexed, processed, queried and searched like a database; and interchanged with other projects, scholars, software, and encoding systems.

The capability to create multiple outputs from a single data source creates several possibilities not available with traditional publishing methods—partly due to the fact that electronic publication also allows near-unlimited space for parallel versions. Using a single data source behind all of these versions, generated dynamically or at least programmatically by XSL Transformations, also reduces the risk of errors being introduced by a human repeating content and trying to keep different versions synchronized; any change, correction, or update only needs to be entered once in the master data source, and all of the versions, indices, and other instances of this data automatically update to reflect the change. A website that publishes a set of inscriptions (or papyri) from a single XML source rendered using XSLT might contain the usual transcription with accompanying commentary, metadata, and images presented for academic use in HTML.

Using a slightly different set of stylesheets, one might present the same data in PDF for more convenient printing, or even a whole typeset volume that can be ordered and printed on demand. The standard, interpretive Leiden transcriptions may be augmented with parallel diplomatic versions (showing text in majuscule case and without word-breaks, editorial corrections or restorations, for example), perhaps on a different page or appearing in a pop-up window. Slightly different versions might also be designed for students, with more emphasis given to the translation and historical commentary, for instance, less to the technical details in the apparatus. Stylesheets may also transform the same text into an audio version for the visually impaired, or render the data and formatting in ways geared toward those with other disabilities. As mentioned above, the indices, tables of contents, concordances, and other summary appendices are generated from the same data via the same technology of XSLT. Because the data in all of these cases is the same, the epigraphist need only do the intellectual work of compiling her publication once, and the content will always reflect this master version of the work.

In other words, as well as offering a choice of rendering styles, this semantic markup also allows us to perform intelligent processes and searches upon the marked-up text. We can perform a word search only for certain words when they are complete and not made up in part or in full by editorial supplements, for example. Certain types of damage may be significant in their own right, such as erasure, which might represent *damnatio memoriae* or the replacement of one name or expression with another. The detailed markup in EpiDoc XML has the capacity to contain this information which would be relegated to a text note or apparatus in a traditional publication, and so not available to a search process or indexing tool.

Project Needs

As discussed above, it is a central goal of the EpiDoc Collaborative to create freely available tools, well-documented advice, and a lively community of training and assistance for EpiDoc projects. We recognize the importance of minimizing the overhead required for a new project to begin using EpiDoc XML to record and publish its inscriptions, or for an existing corpus or database to be converted to the XML interchange format. Some of the early EpiDoc projects such as *Inscriptions of Aphrodisias* spent large amounts of grant money on development, keying of Greek and XML, technical support and programming, and communication with the larger EpiDoc community. This funding is not available to every project, and it would be unreasonable to expect an equivalent investment from all epigraphic and papyrological projects wanting to work in this way. It should be noted, however, that these projects were pioneers and that, as well as publishing new corpora of inscriptions, they were piloting and helping to develop the EpiDoc protocols, guidelines, tools, and the community itself. It is this initial investment that we hope will make the overheads much more tractable for future projects.

Nevertheless a certain amount of preparation and support is still required for a philologist or epigrapher to produce and publish texts in this way. Although the workload may seem to far outstrip that required to publish an epigraphic corpus in book form, I feel it is important to stress two mitigating factors. Firstly, as an electronic publication, some of the extra effort required is the result of there being no traditional publisher to take on the tasks of typesetting, printing, and dissemination. Secondly, it should be recognized that many of these tasks are in fact those that an author will always perform, but carried out in a different sequence. For example, it clearly takes somewhat longer to mark-up the text in EpiDoc XML than merely to type a transcription with Leiden *sigla*—even if conversion tools are used to automate the vast majority of the work. Even more effort is expended in tagging words, names, and places, perhaps even lemmatizing, marking keywords, and otherwise making the highly structured epigraphic edition machine-readable and actionable. Once this work is done at an early stage of the project, however, certain automatic tasks are possible that save further work at a later stage. XSL transformations may be applied to generate indices and other summaries; rather than generating these indices as a one-off task (manually or otherwise) for the publication stage, the indices can also be used as part of the research process. With the first 100 inscriptions of a 5,000-text corpus tagged, the running index of names is already a useful onomastic and prosopographical tool against which each new person can be compared. The same is true of places, keywords, vocabulary, bibliography, locations, and other pieces of information. The task of collating names is perhaps being carried out sooner than in a traditional project, but there is significant pay-off for this effort.

I shall attempt to enumerate here what I see as some of the needs and commitments that a scholar should take into account when considering an EpiDoc-based publication of an epigraphic or papyrological collection. Most importantly, the scholar should be willing to learn about and engage with the new methodologies and technologies of XML and electronic publication. The principal investigator of any project, even one who has many assistants both academic and technical, surely needs to be *au fait* with the principles and operation of the methodologies underlying the project, at least at a theoretical level and preferably with practical experience also. The semantic distinctions recorded in EpiDoc XML and the processes and transformations used to exploit them are essential to the intellectual undertaking of producing and publishing an epigraphic corpus, and should be engaged with by the scholar at every level. The editor therefore needs to be prepared to learn how XML works and what an EpiDoc-encoded inscription looks like, at least well enough to review the work of other editors or conversion tools. She needs to be willing to engage with the EpiDoc community via online fora, mailing lists, and virtual or physical conferences, to share problems, benefit from the experience of others, and return the lessons learned to the community at large. She also needs to be disposed to using the various online sources of reference and documentation, in particular the SourceForge repository for Open Source projects,[20] and the EpiDoc documentation wikis.[21] For an academic who has mastered Greek and Latin, paleography, ancient history, the conventions of our discipline, and the standard requirements of modern publishing, this should not be a terribly onerous new set of skills to learn.

There are few papyrologists or epigraphers, however, who will yet be able to undertake a complete EpiDoc project without some degree of institutional and technical support. At the least, the project needs to have access to: somebody with web design and authoring skills, somebody able to adapt the generic EpiDoc XSL stylesheets to the specific needs of the project, and somebody able to adapt the standard EpiDoc conversion tools such as CHET-C and the Crosswalker for use on project materials. (These skills may all be instantiated in a single person, perhaps even a classically trained research assistant, but I suspect will more often be the combined expertise of a technical support center or research unit.)

I shall briefly address these skills in their likely chronological order, and then consider some of the ways in which the EpiDoc community and tools attempt to help with these tasks. Even if a project researcher has the skills required to mark up texts in XML by hand, almost any collection of more than a few dozen inscriptions or papyri would want to be converted programmatically from a more human-readable (and -writable) form, such as word-processor documents or database records. EpiDoc tools such as CHET-C and the Crosswalker are designed for this purpose, but may very likely need to be customized, finetuned, or extended to suit the idiosyncratic needs of any author's files. A programmer with skills in JavaScript, XSLT, and preferably a high-level language with good text-processing capabilities (e.g. Perl, Python, or Java) would probably be required to carry out this work, although it would be a relatively small task for such a person (requiring days rather than weeks in most cases).

Some programming support will probably also be needed for the transformation and presentation side of the project. Once the texts are in XML, the project needs tools to process and render these files. Although the standard EpiDoc XSL Transformation stylesheets do a good job of rendering XML as Leiden-compliant HTML, and example stylesheets are also provided that produce indices of some of the more common types of keywords, some adaptation or at least customization of these stylesheets will certainly be required. XSL is a simpler script to learn than many programming languages, and is accessible to almost anyone with XML skills, but a better job will be done more easily by someone who is an expert in XSL than by someone who has picked it up rather recently. Depending on the amount of customization required, the number of indices desired, and the complexity of any search tools that are envisaged for the publication, this element may require a fair amount of work for an XSL specialist.

Secondly, in order to publish the results of such a project, the HTML and other electronic output of the XSL Transformations will need to be incorporated into a coherent web design. This is not an especially difficult or time-consuming task in itself, but again a web designer would clearly do a better job than an academic with some amateur HTML-authoring skills.

I should reiterate, however, that the author and researcher on the project obviously needs to be able to make decisions about the exact content, structure, and behavior of the textual material. The central task of designing, authoring, checking, hand-correcting, and perhaps collating the EpiDoc XML files that contain the epigraphic text and editions is therefore one for a trained classical epigrapher or papyrologist, and not for a purely technical assistant. It ought to be evident that only someone with an intimate knowledge of the Greek or Latin texts and ancient history, whether the principal investigator or a hired research associate, has the competence required to make the semantic distinctions in the XML that will engender the sophisticated research output. This means that a classicist responsible for the research needs to be able and willing to understand the principles of EpiDoc XML, to read it (perhaps with technological assistance in the form of rendering tools and research-sensitive indexing), and to write and edit it—even if the majority of the technological burden is carried by conversion tools.

Any project using EpiDoc recommendations to analyze, publish, and/or exchange epigraphic or papyrological texts will almost certainly need to use a combination of the EpiDoc conversion tools and hand-tidying. Individual texts are more easily typed in a word-processor using Leiden, then converted to XML using a tool such as CHET-C, and optionally hand-tidied to fine-tune those distinctions that a machine can not derive from the sometimes ambiguous Leiden conventions. If texts already exist in electronic form in large numbers, a batch-conversion tool will make much lighter work than copying-and-pasting each text individually into the CHET-C interface; whether this is a local customization of CHET-C or the Crosswalker tool will depend on the structure and format of the data. The classicist needs to make decisions about the XML, but certainly does not need to type large quantities of highly structured, near-opaque text full of angular brackets and other codes.

In addition to the tools mentioned above, the EpiDoc community is a repository of experience and advice, made up as it is of many scholars who are performing or have performed projects of this kind using XML. There are many questions involving prospective projects that cannot be answered in this paper: issues such as personnel needs, estimates of costs, time required for certain tasks, and technology issues such as desirable software and hardware. For these sorts of questions, as well as specific technical issues that may arise as a project proceeds, the lively communities on the Markup list and the EpiDoc IRC channel, for example, are likely to be a supportive and enthusiastic source of advice and experience.[22] It is also hoped that the EpiDoc community will be able to continue to offer practical training sessions in the use of EpiDoc XML and tools in the form of week-long summer schools. Such training has already been offered several times in London and Washington DC, with shorter courses in venues as far-flung as Rome and San Diego. Briefer sessions may also be incorporated into epigraphic summer schools for graduate students. Any such courses offered in the future will be announced in the Markup list and other venues.

Final comments

It remains to be seen whether the main potential of the EpiDoc guidelines and technologies will be as a publishing medium for epigraphy and papyrology or as a tool for interchange between existing corpora, databases, schemata, and electronic publications in the field. It is our aim currently, as I believe it has been of the EpiDoc Collaborative from the outset, to work to further both of these outcomes of the technology, and to allow projects that require both scale and depth of markup to coexist and be compatible within the guidelines. If a collection of texts is published using these technologies and following the EpiDoc guidelines for interchange, then not only are the various advantages of XML publication to be had, but if the source XML is also made available, then compatibility with and integration into the large classical text databases and collections will also be facilitated.

Notes

[1] See van Groningen 1932; Wilcken 1932.

[2] E.g. Robert 1983, 9-11 on 'Signes critiques du corpus et édition', where he rejects many of the disambiguating *sigla* introduced by Leiden.

[3] Most notably Dow 1969 and Panciera 1991.

[4] E.g. *SEG* 51 (2001 [pub. 2005]), p. xxxiv; *POxy* 70 (2006), p. xi; *AnEp* (2002

[pub. 2005]), p. 15.

[5] See Nicholas 2000-2004.

[6] Alföldy 1986-2009; Panciera 1999-2009; Carletti/Felle 2003-2009; the confederated database is now accessible via the Electronic Archive of Greek and Latin Epigraphy: http://www.eagle-eagle.it/.

[7] See the report by Panciera 2002.

[8] Elliott 2000-2009.

[9] In particular the Markup list, which can be browsed or joined from http://lsv.uky.edu/archives/markup.html.

[10] Reynolds/Roueché/Bodard 2007; the first publication of the project was Roueché 2004. See also Bodard 2008 and Cayless/Roueché 2009.

[11] World Wide Web Consortium 1996-2009.

[12] Text Encoding Initiative 2001-2009.

[13] SourceForge, available: http://sourceforge.net/ is the home of many software and encoding projects and communities; the GNU General Public License (GPL) is described at http://www.gnu.org/copyleft/gpl.html.

[14] Elliott 2006.

[15] Au/Bodard/Elliott 2007-2009.

[16] Cayless 2006; since this chapter was written, work has begun on a more sophisticated tool for the conversion and tagless editing of EpiDoc texts by the Mellon-funded Integrating Digital Papyrology project; see <http://idp.atlantides.org>.

[17] Elliott/Cayless/Bodard 2007.

[18] Roueché 2004.

[19] TEI Guidelines *s.v.* ‘lb’, <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-lb.html>: “The type attribute may be used to characterize the linebreak in any respect, for example as word-breaking or not.”

[20] The EpiDoc web page at http://epidoc.sourceforge.net/ is the first port of call and gateway to all other information; the source code for the EpiDoc guidelines, tools, and other outputs can be accessed and downloaded from http://sourceforge.net/projects/epidoc/.

[21] The currently active wikis are: the EpiDoc roadmap (http://epidocroadmap.pbwiki.com/); the EpiDoc documentation wiki (http://epidocumentation.pbwiki.com/); and the Inscriptions of Aphrodisias documentation wiki (http://insaphdocumentation.pbwiki.com/). As these sites all reflect projects in development, it is probably best to consider the main EpiDoc web site at http://epidoc.sourceforge.net/ as the most stable address for reference.

[22] For information on the Markup list, the EpiDoc IRC channel, and all other sources of advice and community, see the EpiDoc website at http://epidoc.sourceforge.net/.

References

Géza Alföldy *et al.* (1986-2009), *Epigraphische Datenbank Heidelberg*, Heidelberger Akademie der Wissenschaften, available `www.epigraphische-datenbank-heidelberg.de/` [April 2009].

Zaneta Au, Gabriel Bodard, Tom Elliott, *et al.* (2007-2009), 'The EpiDoc Example Stylesheets', available: `epidoc.sourceforge.net/resources.shtml#xslt` [April 2009].

Gabriel Bodard (2008), ' The *Inscriptions of Aphrodisias* as Electronic Publication: a user's perspective and a proposed paradigm', in edd. G. Bodard & S. Mahony, *"Though much is taken, much abides": Recovering antiquity through innovative digital methodologies*, *Digital Medievalist* 4, available: `www.digitalmedievalist.org/journal/4/bodard/` [April 2009].

Carlo Carletti, Antonio Felle, *et al.* (2003-2006), *Epigraphic Database Bari,* Università degli Studi di Bari, available `www.edb.uniba.it/` [April 2009].

Hugh Cayless (2006), 'The Chapel Hill Electronic Text Converter', available: `epidoc.sourceforge.net/resources.shtml#chetc` [January 2007].

Hugh Cayless, Charlotte Roueché, *et al.* (2009), 'Epigraphy in 2017', in edd. G. Crane & M. Terras, *Changing the Center of Gravity*, *Digital Humanities Quarterly* 3.1, available: `digitalhumanities.org/dhq/vol/003/1/000030.html` [April 2009].

Sterling Dow (1969), *Conventions in editing: a suggested reformulation of the Leiden System*, Greek, Roman and Byzantine Studies Scholarly Aids 2, Durham.

Tom Elliott *et al.* (2000-2009), *The EpiDoc Collaborative for Epigraphic Documents in TEI XML*, available: `epidoc.sourceforge.net/` [April 2009].

Tom Elliott *et al.* (2006), 'The EpiDoc Cocoon Web Application', available: `epidoc.sourceforge.net/resources.shtml#webapp` [January 2007].

Tom Elliott, Hugh Cayless, Gabriel Bodard (2006), 'Crosswalker: EpiDoc Exporter Tool', available: `epidoc.sourceforge.net/resources.shtml#crosswalker` [January 2007].

B. A. van Groningen (1932), 'Projet d'unification des systèmes de signes critiques', *Chronique d'Égypte* 7, pp. 262-269.

Nick Nicholas *et al.* (2004), *Beta Code Manual*, Thesaurus Linguae Graecae website, available: `www.tlg.uci.edu/BCM2004.pdf` [January 2007].

Silvio Panciera (1991), 'Struttura dei supplementi e segni diacritici dieci anni dopo" in *SupIt* 8, 9-21.

Silvio Panciera (1999-2009), *Epigraphic Database Roma*, Università di Roma I, available `www.edr-edr.it/` [April 2009].

Silvio Panciera (2002), 'Commissione "Épigraphie et Informatique" de l'AIEGL: Relazione (1997-2002) nell'Assemblea Generale di Barcellona (6 settembre 2002)', available `www.edr-edr.it/Documenti/Document1_it.html` [January 2007].

Joyce Reynolds, Charlotte Roueché, Gabriel Bodard (2007), *Inscriptions of Aphrodisias*, available: `insaph.kcl.ac.uk/iaph2007/` [April 2009].

Louis Robert with Jeanne Robert (1983), *Fouilles d'Amyzon en Carie*, Paris: De Boccard.

Charlotte Roueché *et al.* (2004), *Aphrodisias in Late Antiquity: The Late Roman and Byzantine Inscriptions*, revised second edition, available: `insaph.kcl.ac.uk/ala2004/` [April 2009]. (A digital second edition of the 1989 title: *Aphrodisias in Late Antiquity*, Society for the Promotion of Roman Studies Monograph 5, London.)

Text Encoding Initiative (2001-2009), *TEI Guidelines*, available: `www.tei-c.org/` [April 2009].

Ulrich Wilcken (1932), 'Das Leydener Klammersystem', *Archiv für Papyrusforschung* 10, pp. 211-212.

World Wide Web Consortium (1996-2009), Extensible Markup Language (XML), available: `www.w3.org/XML/` [April 2009].