Journal of the Text Encoding Initiative

Issue 7 | 2014
Reaching out, Opting in

Towards an Interoperable Digital Scholarly Edition

Desmond Schmidt

Publisher
TEI Consortium

Electronic version
URL: http://jtei.revues.org/979
DOI: 10.4000/jtei.979
ISSN: 2162-5603

Electronic reference
Desmond Schmidt, « Towards an Interoperable Digital Scholarly Edition », *Journal of the Text Encoding Initiative* [Online], Issue 7 | November 2014, Online since 01 January 2014, connection on 02 October 2016. URL : http://jtei.revues.org/979 ; DOI : 10.4000/jtei.979

Journal of the Text Encoding Initiative

Issue 7 | 2014
Reaching out, Opting in

Towards an Interoperable Digital Scholarly Edition

Desmond Schmidt

Publisher

Electronic version

Electronic reference

Towards an Interoperable Digital Scholarly Edition

Desmond Schmidt

1. Introduction

1 The Text Encoding Initiative has sought to define a "standard or normalized practice" (Ide et al. 1988) for the encoding of a variety of text types since 1988, focusing on a uniform encoding format (SGML, later XML) and a recommended set of tag and attribute names. Although the purpose of the TEI Guidelines (TEI Consortium 2014) is to provide a general encoding scheme for texts of all types, its main applications to date have been in the field of historical literary and documentary texts. Of the 158 projects listed as using the Guidelines on the TEI website,[1] 123 fall into this category. TEI-encoded texts thus often form an important part of a digital scholarly edition (DSE), which may be defined as the modeling in the digital medium of the scholar's interactions with the text. This naturally includes much more than mere transcriptions: exegetic commentary, textual notes, contextual information in the form of biographies and other personal data, facsimile images of manuscripts and books, as well as functionalities such as the ability to annotate, select versions, and compare transcribed text with the originals (Gabler 2010). But the encoded transcriptions of the original sources form the crux of it: they are the things upon which all this depends.

2 The change in context between the pre-Web world of 1988, its isolated microcomputers and CD-ROMs, and the modern connected, mobile world of Web 2.0 is stark. Texts now have a different purpose: they need to be much more than simply exchangeable via disk or email, they need to instantly respond to heterogenous needs. Inevitably, this has resulted in a growing discontent with traditional approaches to encoding one-off literary and historical documentary texts (Mueller 2013; Robinson 2010). This discontent has focused on the size of the TEI Guidelines (currently 553 tags), the consequent difficulty of comprehending it enough to use it (Usdin 2009), and its inability to elegantly represent

overlapping structures that are common in born-material texts (Renear, Mylonas, and Durand 1993). But arguably the most serious problem, now generally recognized (Unsworth 2011; Bauman 2011), is that TEI-encoded texts are not interoperable: that is, they cannot be fully used in various applications without preliminary, and often substantial, conversion.

3 The Web 2.0 world takes such interoperability for granted. No one is surprised by Word documents that load (almost) flawlessly into Google Docs, or by the possibility of editing image files produced in various formats by others, or by the mundane (but still amazing) fact that every Web page is readable in a variety of browsers, on a variety of devices, or that the complex interactivity of those pages works in each such environment. But all of these things did not happen by chance; they had to be engineered to work that way.

4 These forces of change have also affected the digital humanities. It has been argued that much of the recent growth in the field is precisely due to the increasing use of social networks and in the rise of mobile and interactive use of data at the expense of older static forms (Jones 2014). Recent developments in the construction of international and national repositories of humanities texts (TextGrid, HuNI, INKE, and Islandora),[2] experiments in crowd-sourcing (Causer, Tonra, and Wallace 2012), and "social" applications designed to construct DSEs by contributions from geographically dispersed scholars (Crompton and Siemens 2012), have all underlined the importance of interoperability. Another area where digital scholarly editions need to work better together is in providing a stable means for parts of editions to be referred to (Blackwell and Smith 2012), and for the software supporting those references to interact with various applications over time.

5 So what can be done to make these digital "surrogates" of culturally important texts more generally amenable to scholarly processes? How can software be built that successfully models the former interactions of scholars with the print edition: the ability to compare, annotate, find, reference, catalog, edit, and study texts? And how can these operations be shared effectively? If all these processes are now to be remodeled in the digital medium, the thing on which they operate, like the printed book before it, must itself be an interoperable object. One way to achieve this may be to separate out the different kinds of information currently stored in the same file format. TEI-encoded data falls naturally into four classes: markup, annotation, metadata, and plain text versions. It is no accident that these four classes correspond to the forms of data expected by modern applications. For example, metadata is now mostly stored separately from the texts it describes in repositories such as Fedora[3] or Dspace,[4] not embedded in a TEI header. Annotations are also external, as in the Open Annotation Data Model,[5] not embedded in the same text they describe. Markup systems like XML and SGML were originally designed to format linear text (Goldfarb 1996), not to represent variations in the very information they are describing. Separating out markup from text and splitting texts into their variant components would thus clarify their respective roles in the digital scholarly edition. Plain text is also the most widely supported format for text analysis. Of the 54 current text analysis tools listed on the DiRT Wiki,[6] 49 can read plain text.

6 The rest of this paper treats each of these points in more detail. Section 2 describes the nature of the interoperability problem. Section 3 looks at how and why markup is best stored separately from the text it describes. Section 4 looks at metadata and annotation, and how it also benefits the overall scheme of "big tent" humanities to record it outside of transcriptions. Section 5 looks at how texts containing internal variations can be split

into individual coherent layers. Section 6 outlines how the proposed reorganization of the data of the DSE could be expressed as a bundle of highly interoperable resources. The conclusion then draws together each of the points and makes the overall case for the benefits of this approach.

2. Interoperability and the TEI Guidelines

7 *Interoperability* may be defined as the property of data that allows it to be loaded *unmodified* and fully used in a *variety* of software applications. *Interchange* is basically the same property that applies after a preliminary conversion of the data (Bauman 2011; Unsworth 2011), and implies some loss of information in the process. Interchange can thus be seen as an easier, less stringent, or less useful kind of information exchange than pure interoperability.

8 From its inception, the TEI Guidelines were conceived as a format for interchange, specifically to overcome the then prevalent use of multiple character encoding standards in 1987. But it is clear from the original grant proposal (Ide et al. 1988), and from the Poughkeepsie principles drawn up after the inaugural TEI conference in 1987 (TEI Consortium 1988), that the TEI was also conceived as an interchange format based on standardized tags that would eventually form the basis of an interoperable format for the writing of *shared* software:

> If a common encoding scheme existed ... the materials created by these projects would be in a uniform format ... Even more important, we can assume that the existence of a common format will prompt software developers to accommodate this format.... Therefore, the materials created by projects over the next decade could serve as input to as-yet undeveloped software designed for any number of text analytic tasks. If both the creators of textual scholarly materials and software developers utilize a common encoding format, the texts may be used with any software package.
> (Ide et al. 1988, 1.4)

9 Whatever this meant at the time, Syd Bauman (2011), one of the original editors of TEI P5, has since observed that interoperability of TEI-encoded texts today—that is, the exchange of unmodified TEI files between different programs—is "impossible." Bauman is more optimistic about interchange, but he admits that interchange remains "difficult" and that it involves considerable human intervention. Martin Mueller, as chairman of the TEI Board, remarked that in his experience, detailed TEI markup is generally ignored in practical applications, and currently offers no advantage over plain text, HTML or ePub formats (2011, 5). A recently published report on Project Bamboo likewise described the stripping out of all TEI encoding so that texts gathered from various sources could be made to interoperate (Dombrowski and Denbo 2013). As John Unsworth (2011) says: "The *I* in TEI sometimes stands for interchange, but it never stands for interoperability."

10 These express admissions by those involved closely with the TEI seem at odds with the user community's needs for standardization and interoperability that gave rise to the TEI in the first place (Cummings 2007, §4). The interpretive data contained in the tags is instead locked up in specific projects or "digital islands" (Robinson 2010, 158) that are of little use to the wider community of scholars. As Martin Mueller asks: "What about the added value of TEI specific encoding for the historian, linguist, philosopher, literary critic etc.? How can they decode or get at it, and what does it do for them? The answer is that for the most part they cannot get at it at all." (2011, 5)

11 To those involved in the creation of digital scholarly editions, interoperability seems to matter a great deal. The whole point of the large European Interedition project was to create "an interoperable supranational infrastructure for digital editions."[7] Likewise, one of Peter Robinson's five desiderata for scholarly digital editions was that "all the materials in a digital edition should be available independent of any one interface" (Robinson 2013). In a similar vein Dot Porter writes: "Notably, from 1992 through today, the papers, sessions, and workshops at the Congress that focus on digital editing focus on the creation of those editions, but there is very little if anything to be found on how those editions might be used by the scholarly community" (Porter 2013). Or as Maria Morrás puts it: "I doubt much that there will ever exist definitive computer programs, or even a stable format for presenting and connecting texts in hypermedia archives.... It is essential that texts are transcribed in accordance with a universal system which enables the transference from one program to another ..." (Morrás 2003, 235—my translation). Indeed, if sharing of edition data stored in repositories, and collaboration on their preparation by contributors across the world, is to happen at all, it seems clear that the required benchmark is data *interoperability*, not mere data interchange. As Martin Mueller complains, "the TEI keeps insisting on a distinction between *interoperability* and *interchange* that makes very little sense to folks outside the TEI's discursive realm" (2013).

2.1 The TEI Paradox

12 And here is the paradox for the modern digital humanist who works with transcriptions of the textual content of original artifacts. Everyone knows that XML itself, the base technology for the TEI Guidelines, is a highly interoperable format. No one is disputing that there are many applications able to understand XML, such as Apache Cocoon, XML parsers, XQuery, oXygen, XSLT processors, etc. Valid XML data can be loaded unmodified into such applications and can be successfully parsed, merged, edited, searched, and transformed. But in spite of these properties of XML itself, information encoded at the TEI level, at the "tag" level, is mostly *not* interoperable, for reasons first explained most cogently by Alan Renear (2000). Renear argued that there is an important distinction to be drawn between

1. a title tag inserted by the transcriber of a physical document as an interpretation of what he/she sees (but it may not be true), and
2. the same tag used by an author to declare that his/her digital text is in fact a title.

TEI tags are usually of type 1, because they are the result of human interpretation. Most other XML tags, however, are of type 2.

13 Renear characterizes the distinction as one of mood. Type 1 corresponds to the indicative mood, and type 2 to the imperative, where the author of the text is effectively issuing the command: "be a title!" Markup is so often created in situation 2 that one can easily be tricked into thinking that it is always so. For example, markup generated by a machine is easy for another machine to read. All that is needed is to reverse the algorithm that wrote it. An example using the TEI Guidelines would be an XML text generated by a natural language parser. Such XML files *may* be interoperable, if standardized, because they are part of a machine-to-machine conversation. Even a program that translates objects created by humans using a GUI interface, such as a drawing program or a word-processor, can save the result in an interoperable format because it is the *program* that reads and writes the tags, not the human. An example of such a format is SVG (Scalable Vector Graphics), which is an XML format. Drawings saved in SVG[8] can be rendered and edited in

a variety of programs. Even when a human creates a file in HTML the result is usually interoperable because all the tags created are of type 2. But when a humanist transcribes original historical documents, such as the contents of a printed novel, the result is markup that is *not* interoperable because what is guiding the selection and application of the tags is a human brain, and every step the human takes is an interpretation of type 1.

14 Patrick Durusau illustrates the significance of this distinction with a practical example. He enumerates more than *four million ways* to encode a simple sentence taken from a printed book using the TEI Guidelines. And this huge figure "is by no means exhaustive." This illustrates, as he says, "the unlikelihood that any two encoders or even the same encoder on different days will make, without formal guidance, the same decisions" (Durusau 2006, 302). Durusau argues that this demonstrates the importance of documenting encoding choices, of modeling texts in advance, of training encoders, checking syntax, and reviewing transcriptions. Although all these measures taken together will probably prove effective, the author knows from the experience of having worked for more than ten years on the Vienna Edition of Wittgenstein that these methods are very costly to implement, and in cases where the contributors to a project are from different geographical locations, they are probably also hard to enforce.

15 The reason behind this difficulty is the nature of being human: everyone sees different features when they look at the same text in a manuscript. And everyone understands the meaning of the tags to which they must map those features differently. It is easy to make texts syntactically correct, even conform grammatically to a given markup scheme, but one also must ask, are they internally *consistent*, and can they be *kept* in that state in the face of continued editing? As Tommie Usdin (2002) points out, having more than one way to encode the same thing increases choice but does not make things any easier; in fact it *magnifies* the work of encoding.

2.2 Some Examples

16 If Renear is right, then it is the *illocutionary force* of TEI markup as applied to original documents that leads to its non-interoperability, and not simply the size of the tag set. Even if the TEI Guidelines were reduced to a single tag, `<title>`, there would still be dispute as to which pieces of text were titles and which were not:

> Suppose *The Babylonian Captivity of the Church* does not display as a title. Thinking that the stylesheet author has failed in some regard, the user attempts to search the document for titles of works cited. Oddly enough far fewer titles appear than are known (or assumed) to occur in the text.
> (Durusau 2006, 302)

17 Although entirely plausible, Durusau's example is still hypothetical. However, in the DTA (Deutsches Textarchiv), an attempt to homogenize a number of digitized texts encoded in various subsets of TEI led to this same scenario being played out for real:

> machine-exploitable extraction of document components such as 'retrieve all letters of the document collection' or 'display all quotations in a chapter' pose an enormous problem since division types or entity encoding for quotes do not have to be realized in an ubiquitous way across document collections
> (Geyken, Haaf, and Wiegand 2012, 384)

18 Another real-world example is the interesting experiment undertaken by Dombrowski and Denbo (2013) as part of Project Bamboo. Their idea was to develop an "XSLT web service engine to transform XML-marked up bibliographic entries into HTML." Unlike

markup for literary documents, the TEI tags for marking up bibliographies are relatively constrained. The project was planned to take three weeks. In the end it took a year and involved a considerable amount of manual data and stylesheet manipulation:

> This revealed markup inconsistencies within the TEI-encoded data (such as the ordering of authors' first and last names), which Hooper then revised. The revised TEI, in turn, brought to light legitimate variation in the data that the XSLT stylesheets did not correctly account for. In this way, both the data and the stylesheets underwent iterative development for a number of months.
> (Dombrowski and Denbo 2013)

This demonstrates the difference between interchange and interoperability. If the files had been in an interoperable format in the first place they could have been ingested and processed immediately.

19 Another, more subtle example is the experiment in interpretative encoding conducted by Kate Singer (2013). When her students marked up the same poem using TEI tags for different tropes, each group of students ended up recording very different features in the same text, and at the end of the experiment it became clear that all their work had been frozen in their individual copies:

> my students began to dream of an interface that allowed them to bring our class's particular collections of text and commentary to bear on a primary text, one with the ability to then permit future classes to render their own versions in the space of the same edition.
> (Singer 2013)

The poetic tropes recorded by embedding markup tags in the texts they described had led to a form of non-interoperability: it prevented them from sharing their interpretations with other future students, or from collaborating with other current groups.

2.3 Encoding Documents for Different Purposes

20 If it is hard to make files all with the same purpose interoperate, how far-fetched is it then to expect that files with different purposes might also one day interoperate? For example, the TEI website[9] lists a number of projects creating linguistic corpora of historical texts, like the Base de Français Médiéval, which contains 4.7 million words, and has been morphologically tagged, but the material is also of considerable literary and historical interest.[10] Another example is the Croatian Language Repository,[11] which contains a linguistic corpus that includes novels, short stories, drama, and poetry from the mid-nineteenth century onwards. A similar picture emerges from the Icelandic Online Dictionary and Readings. This linguistic corpus consists of readings taken from newspapers and both modern and nineteenth-century literature.[12] Yet another case is the syntactic tagging of the fourteenth-century French work by Jehan de Joinville, entitled *La Vie de Saint Louis*, a mixture of standard and Champenois French, which is "extremely important for historical and literary, as well as for linguistic reasons" (Estival and Nicholas 1997). One may well ask the question, what is the point of digitizing all these texts if the end result will not be reusable for a variety of purposes? Repurposing a text that has already been marked up for one application should not mean that it must be re-digitized or undergo a time-consuming conversion, possibly damaging the underlying content in the process. The tags for linguistic corpora have little in common with those used for literary texts, but often the underlying material is shared between them. The next few sections will examine how and why digitized texts should be prepared from the start with such re-use already in mind.

3. Removing Markup from the Text

21 One obvious remedy to this problem is to remove the main source of non-interoperability, namely the embedded markup itself, from the text. By removing it, the part which contains all the significant interpretation can later be added or substituted at will.

22 What remains when the markup is removed is a residue of plain text that is highly interoperable, which can be exchanged with other researchers, just as the files on Gutenberg.org are downloaded by the tens of thousands every day (Leibert 2008). However, if one suggests this to someone who regularly uses TEI-XML, the immediate objection is made that this will solve nothing, because even plain ASCII texts are still an interpretation of what the transcriber sees on the page (e.g. Sperberg-McQueen 1991, 35). This point, although valid to a degree, misses an important distinction.

23 But first consider what exactly is the nature of the interpretation exercised when a text is transcribed. A digital text encoding system, even one as expansive as Unicode, has a limited power to capture the features of abstract "text," which may be understood as the entire content of a page to be transcribed (Sahle 2013, 244). Certainly there is much that cannot be represented directly in this way: for example, unconventional characters found in manuscripts and early printed books, including abbreviations.[13] Many of these cases are ligatures, which can be encoded in their decomposed forms. Unicode is "a *character* encoding standard, and is not intended to standardize ligatures or other presentation forms" (Unicode 2010). For example, there is no Unicode character for old Latin *sescuncia* (like a pound-sign, means "one eighth"), since it can be composed from *semuncia* (character 10192) and an EN-dash (Perry 2006, 4). However, even in modern texts there are inline mathematics and graphics, which pose a similar problem. In such cases markup must be used to extend the capabilities of plain text. But let these exceptions be set aside for the moment, because the discussion on markup below also applies to them. Encodable character data or "plain text" is so overwhelmingly prevalent in literary and historical documents that it should be treated as a separate case.

24 An encodable character on a page is either illegible, unclear, or clear. If the character is simply illegible, then a gap, qualified by some markup describing how long it is and how it has been rendered illegible can be added (TEI Consortium 2014, 3.4.3). If the letter is not clear then it admits of several various readings. It is thus like a variant and can be treated by the mechanisms for recording variants. But in by far the most common case, if the character symbol is clear and the letter has, say, the shape of the letter *t*, choosing to encode it as the Unicode character code 116 for *t* is probably not an interpretation, because no reasonable person would dispute the point. On the other hand, choosing to record it or not *is* an interpretation. Patrick Sahle (2013, 244f) argues in his "pluralistic text model" that different texts will be needed for different purposes. For instance, whether or not to encode the running header, or the index at the back, or the text in an image, would be an editorial decision. This interpretation is binary in nature: the editor can only choose to encode it or not.

25 On the other hand, the interpretation about which format or logical structure an italicized *t* belongs to involves assigning it to one of a myriad of possible encodings. In the TEI-Lite schema alone one may choose, for example, between `<hi rend="italic">`, `<hi rend="italics">`, `<foreign>`, `<stage>`, `<term>`, `<soCalled>`, and

<head>. The problem with choosing one formulation of "italics" or another and inserting it into the text is that this constitutes a declaration of intent to write software that will act upon those specific interpretative codes. Even just encoding it in any kind of XML declares an intent to process it in an XML-aware application. Encoding it in a specific customization of TEI-XML presupposes that there is software that can understand that encoding, including its interpretation of the TEI Guidelines, which specifies how the chosen attributes and tags should be applied to the text in question.

26 Encoding a text in Unicode is also admittedly a declaration of intent to process it in a Unicode-aware application, but programs that understand the Unicode character encoding are now practically universal.

27 To achieve interoperability—the ability to load a transcription into various programs without modification—it is thus obvious that the markup must first be removed from the text, because it is the markup that contains virtually all of the interpretation, and what little interpretation remains does not stop the text from being interoperable.

28 Admittedly, there are many features recorded in markup that are really part of the "text" and vice-versa, such as tabs, spaces, and carriage-returns—which are, at least in part, layout instructions found in plain text—and paragraph breaks, which are a kind of text content described via markup. However, since interoperability is a technological property (the ability to load a file unmodified into several programs), the distinction must be drawn on technological grounds. There are simply too many ways to mark things that might be interpreted as paragraph breaks—for example, , <ab>, <sp>,—for any combination of text plus basic markup to serve as an interoperable format.

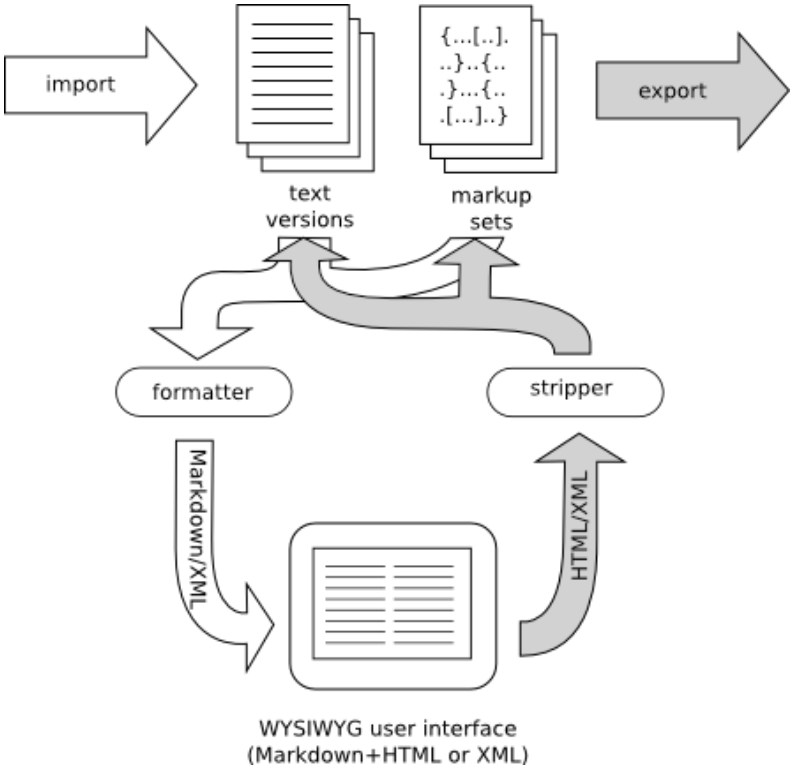
3.1 A Stand-off Editor

29 So how would this work in practice? Figure 1 sketches the design of a practical web-based editing system that could be assembled using existing components. At the bottom the user interface is just a standard WYSIWYG editing environment for XML or a Markdown-like language (adapted for TEI). The latter are now ubiquitous on the Web and have the advantage of offering a forgiving syntax.[14] Rather than parsing the user-input formally, they convert it into approximate HTML, then "tidy" it into valid HTML. The user can thus see the result immediately without receiving any puzzling "syntax error" messages. When the user presses *save* the client sends to the server the HTML or XML (depending on the interface), which is immediately stripped of markup, and the two components are stored separately.

30 Only one editorial layer or internal version of the document can be edited at a time. If there are other stand-off markup sets attached to this version they could be updated on save, by first computing the differences between the submitted copy and the one stored on the server. These differences could then be used to delete, curtail or expand the range of existing markup properties (Nelson 1997; Vulpe and Owens 1998). The user would be entirely unaware that there were any stand-off properties attached to the text; the formatting would appear naturally inline, but the benefits of the separation would always be present in the system. In this form the text+markup can be saved as a coherent copy through export, or reformatted for another purpose using a different markup set or sets. The ability to convert stand-off properties into valid HTML already exists in the formatter tool in AustESE.[15] This could easily be extended to generate Markdown or XML. When the user requests that version again, the digital document is reconstructed from the stand-off

markup properties and the base text and a new XML or Markdown document is created for editing. This data flow has the advantage of silently fixing any coding errors present in the user's input, and painlessly enforcing the correct coding syntax.

Figure 1: Design of a stand-off editor



31 Although this design is not yet fully realized, much of it is. The import and export facilities, from XML to plain text plus markup, the export to a variety of interoperable formats including HTML, XML, and plain text, the formatter and stripper programs, already exist as components of AustESE. All that is lacking is adapting it to an editing interface such as Markdownviewer. But this would also not preclude the use of XML-based WYSIWYG editors like the Islandora TEI editor (Stapelfeldt and Moses 2013) or the Standoff Markup Editor[16] being developed at Loyola.

4. TEI, Metadata, and Annotations

32 The standard definition of metadata as "data about data" seems to pose more questions than it answers. Metadata has come to mean many things, including any form of data that describes digital resources on the Web. But in the context of TEI, metadata means data describing the digital document as a whole, and in TEI it is embodied in the Header, which is an obligatory part of every TEI document. (In this section *document* refers to the digital surrogate, not the physical document.) The problem with embedded metadata is that they are not interoperable. The more common approach now is to store metadata separately from the documents they describe (Haynes 2004, 107). The increasing tendency toward user-driven data models for repository design, inspired by the growth of cultural resources websites, has fueled the development of cross-disciplinary repositories and

increased demand for interoperability of metadata (Spinazzè 2004). This has led to two basic strategies:

1. Federation of existing metadata formats by specifying a translation of metadata properties between formats.
2. Reduction of the metadata to the lowest common denominator across the whole collection. The most common format used for this purpose is Dublin Core,[17] as used in DSpace, or even the more reductionist approach used in TextGrid.

33 Since the advent of "big tent" humanities the digital scholarly edition has had to contemplate a future within a mixture of resources (images, video, audio, or other document types) from related disciplines like archaeology, sociology, history, philosophy, and art history. Functionally, if metadata about a document is not available in a separate database, retrieval is complicated and slow. This means that either the TEI header has to be extracted and transformed into the repository standard metadata, or generated and inserted into the document, based on the repository metadata. Either way, the TEI header looks increasingly as if it is redundant in its present form.

34 The primary uses of metadata are to describe resources and to aid in their discovery (Day 2001). But metadata also play a role in management and description of a resource. Once the resource is found, it would help the user to know more about it. But here the main problem is cost: detailed metadata are expensive to produce. So two forces—cost and the need for interoperability—are both pushing towards external, brief, and standard metadata. The TEIHeader, on the other hand, is embedded, detailed, and non-standard.

35 One quarter of the TEI Guidelines are dedicated to metadata. Part of the bulk comes from the specialized manuscript description module (`msdescription`). This can be part of the TEIHeader, or it can be used as a separate document format. Apart from this module, the TEIHeader itself mostly contains tags that are already provided by other metadata schemas, such as MODS,[18] Dublin Core, and EAD,[19] which is a serviceable substitute for TEIHeader. Removing it altogether, and storing the information externally, would significantly reduce the complexity of the TEI Guidelines without harming its usefulness. Applications that need `msdescription` could simply connect the separate manuscript descriptions to the source texts, using standard metadata.

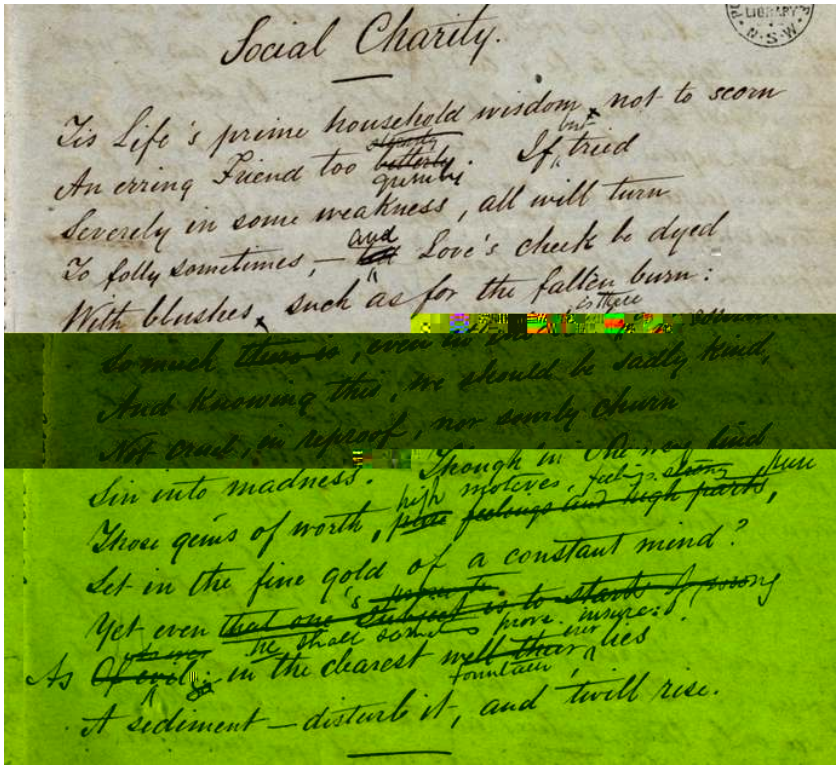
4.1 Annotations

36 Embedded annotations can also be removed from TEI texts. The elements `<note>`, `<interp>`, and `<interpGrp>` describe content that, like metadata, is *about* the text, not the text itself. These are really annotations, and should ideally be represented via the established standards and practices of *external* annotation (Hunter and Gerber 2012). Annotations are stored in *triple stores* or graph databases like Neo4J,[20] which record the identifiers of each component of the annotation and its data. Keeping track of how all these objects interrelate is a specialized task that should be assigned to a dedicated annotation engine. And annotation should point to the document, not the other way around. Otherwise, any alteration to the annotations will break the document. Strangely, the `<interp>` element in TEI does exactly that: it is assigned an `id`, and then the textual element *points to it* via its `@ana` attribute. A similar awkwardness can be seen with `<span>`, which can be used similarly to embed short annotations directly in the text. The problems this caused for Singer's students have already been noted in section 2.2 above. As with metadata, the TEI mechanisms for annotation need to be brought more into line with modern practice.

5. Splitting the Text into Layers

37 Even though markup is subjective, and makes the transcription non-interoperable, it is often said that this doesn't matter, because stripping markup from TEI-encoded texts is "a piece of cake" or "five minutes' work."[21] This section will argue and demonstrate that it is in fact far more trouble than five minutes' work, and that there is no guarantee, even after several weeks of programming, that one may obtain consistent results for TEI files that use any form of variant encoding, including abbreviation and expansion and spelling normalization (i.e. `<mod>`, `<subst>`, `<choice>`, `<sic>`, `<corr>`, `<add>`, `<del>`, `<abbr>`, `<expan>`, `<orig>`, `<reg>`, `<app>`, `<lem>`, and `<rdg>`). Tags are stripped from XML far more often than might be imagined. Whenever a program compares the content (not just the structure) of one XML file with another, whenever it analyzes the text content of a file, what the program sees is effectively the digital document stripped of all its tags. The example chosen for this demonstration is the poem "Social Charity" from manuscript C376 by Charles Harpur, one of the first Australian poets. This documentary version dates to 1851, but there are four others extant, dating from 1848 to 1867. Consider the following half-page of MS C376:[22]

Figure 2: Page 111 of C376 (00000059.jpg)



38 This exhibits only a modest amount of revision—for an autograph. Here is the transcription, as produced by the encoding team, omitting the header for brevity:

Example 1: A transcription of Page 111 of C376

```
<head>Social Charity</head>
<lg type="poem"><l>Tis Life's prime household wisdom<del>,</del> not to
scorn</l>
  <l>An erring Friend too <app><rdg><del>bitterly</del></
rdg><rdg><del>sternly</del></rdg> <rdg>grimly</rdg></app>. If <add>but</
add> tried</l>
  <l>Severely in some weakness, all will turn</l>
  <l>To folly sometimes,-<del>till</del><add>and</add> Love's cheek be
dyed</l>
  <l>With blushes<del>,</del> such as for the fallen burn:</l>
  <l>So much<del> there is</del>, even in the best <add>is there </add>to
mourn.</l>
  <l>And knowing this, we should be sadly kind,</l>
  <l>Not cruel, in reproof, nor sourly churn</l>
  <l>Sin into madness. Though in One we find</l>
  <l>Those gems of worth, <del>pure feelings and high parts</
del><add>high motives, feelings <del>strong</del><add>pure</add></add>,</
l>
  <l>Set in the fine gold of a constant mind?</l>
  <l>Yet even <app><rdg>that one <del>subject is to</del> starts</rdg>
      <rdg><del>that one's prone to starts of wrong</del></rdg>
      <rdg><emph>he</emph> shall sometimes prove insure:</rdg></app></l>
  <l><app><rdg><del>Of evil</del>: in the clearest <del>well <sic>their</
sic><corr>there</corr></del> lies</rdg>
      <rdg><del>As ever</del> in the clearest fountain lies</rdg>
      <rdg><del>So</del> in the clearest fountain ever lies</rdg>
      <rdg>As in the clearest fountain ever lies</rdg></app></l>
  <l>A sediment—disturb it, and ’twill rise.</l></lg>
```

Example 2: A stripped transcription of Page 111 of C376

```
Social Charity
Tis Life's prime household wisdom, not to scorn
An erring Friend too bitterlysternlygrimly.
If but tried Severely in some weakness, all will turn
To folly sometimes,–tilland Love’s cheek be dyed
With blushes, such as for the fallen burn:
So much there is, even in the best is there to mourn.
And knowing this, we should be sadly kind,
Not cruel, in reproof, nor sourly churn
Sin into madness. Though in One we find
Those gems of worth, pure feelings and high partshigh motives, feelings
strongpure,
Set in the fine gold of a constant mind?
Yet even that one subject is to starts
that one’s prone to starts of wrong
he shall sometimes prove insure:
Of evil: in the clearest well theirthere lies
As ever in the clearest fountain lies
So in the clearest fountain ever lies
As in the clearest fountain ever lies
A sediment–disturb it, and ’twill rise.
```

This is likely to be what a text-analysis program will see if it reads this particular XML file. Treatment of both whitespace and alternatives is obviously problematic.

40 As regards whitespace, spaces cannot be placed between all elements, since this is often not desired, as in the cases of partially formatted or canceled words and individual letters or syllables. And in XML not all whitespaces are equal: some are for pretty-printing the XML itself and some are content (Bray et al. 2008, §2.10), and a parser cannot tell which is which without reference to a schema. Furthermore, even if the parser uses a schema it will likely join up alternatives, because there is no significant whitespace between `<rdg>` elements in an `<app>`.

41 As regards searching, the juxtaposition of alternatives in the same text stream will lead to invalid results. In lines 13–15 “that one subject is to starts,” “that one’s prone to starts of wrong” and “he shall sometimes prove insure” belong to different layers, but a search engine would treat them all as part of the same sentence.

42 The solution to all this seems simple enough. One can just write a program that understands this particular encoding, and is able to tease apart the layers, or to take the first or last layer and discard the rest. This splitting into layers is an accepted technique, as used in Gabler’s edition of *Ulysses* (Gabler, Steppe, and Melchior 1984, x), and also in HNML (Zapf 2006). But how easy is it, exactly?

43 Consider the sequences of variants in line 3 and in lines 16–19 in example 1. In line 3 there are three successive alternatives, and in lines 16–19 there are four. Since the last alternative in line 3 is uncanceled it belongs by default to both the third and fourth layers of lines 16–19. It is “by default” because in autographs like this there is usually no way of telling which layer in line 3 was current when the unrelated changes were made to lines

16–19 (Pierazzo 2009, 185). By marking it up as a sequence of corrections, whether as <rdg> or via <add> and <del>, there is already an implicit ordering of variants. The only difference is that here the alternatives are being assigned to default numbered layers; otherwise the information recorded is exactly the same. (A diplomatic rendition could still be produced from the layered representation.) So the main objective is to read coherent layers from the marked-up text, but for this to work, alternatives in one place must be coupled with sensible alternatives elsewhere. Table 1 shows how the alternatives are linked by the splitter program in AustESE, which is used for importing TEI-XML files:

Table 1: Linking of alternatives by the splitter program in AustESE

Layer 1-sic	bitterly	...	Of evil: in the clearest well their lies
Layer 1-corr	bitterly	...	Of evil: in the clearest well there lies
Layer 2	sternly	...	As ever in the clearest fountain lies
Layer 3	grimly	...	So in the clearest fountain ever lies
Layer 4	grimly	...	As in the clearest fountain ever lies

44 Since the <sic>/<corr> is an editorial directive, rather than an authorial change, it effectively splits layer 1 into two sublayers. Writing a general program to extract such layers from any TEI file is hard because different encoders use different ways to record deletions and alternatives. For example, the encoders here opted to mark text inside <rdg> codes that disappeared in the next layer with <del>, treating it effectively as a crossing-out format. But the <del> codes outside of the <app> *do* create new layers: that is, they are not mere formats. Another encoder might find that too confusing, and regard all <del> codes as introducing a new layer, forcing the program to be adjusted. And what if "grimly" had been canceled instead of "sternly?" Then "sternly" would replace "grimly" in layer 4, but not in 3. And what about the bare <add> in line 3: the content of this element belongs to layers 2, 3, and 4 but not to 1.

45 It is obvious from considerations like these that what seems on the surface to be a simple problem is actually very messy and difficult to handle. A program has to compute all this correctly, for as wide a range of texts as possible. In fact it took the author a week of continuous work to adapt an already existing splitter program to work with this material.

46 These problems would all disappear if layers were stored separately. This would greatly simplify editing, since the user and the programmer would only have to deal with one layer at a time. The simple text editor described above in section 3 could be used, since all the complex markup would already be expressed through layers.

6. Putting it All Together

47 As argued here, one way forward is to divide data into functional categories. A digital scholarly edition could be re-expressed as a bundle containing:

1. plain text or HTML versions, one per internal layer, for each document that witnesses the work in question
2. separate markup if plain text is used
3. annotations, and

4. metadata about the documents stored separately from the text.

48 AustESE currently uses a zipped folder structure to represent all this information, including paratextual information such as biographies, in as application-independent a manner as possible. Alternative formats like HTML, TEXT, MVD, and XML are provided for the source documents, and the application that is reading it only has to choose one. Single or multiple editions can be stored in the one container, and these can be uploaded to or downloaded from digital scholarly editions on the Web using the psef-tool,[24] to create a portable scholarly edition format. In future psef-archives may be expressed in standard formats such as EPUB.[25] However, EPUB3 doesn't currently support annotations, although they are likely to be added to the next version of EPUB.[26]

49 But the main advantage of this entire approach is that, however it is realized, the model is designed to support interoperability from the start. Scholars need a separate package of data that they can exchange and use in a variety of programs and platforms, a package they can identify with clearly as a "digital scholarly edition," however it is rendered or edited in practice.

7. Conclusion

50 If there is one point to be drawn from this whole question of interoperability, it is that what is being advocated here is a divide-and-conquer approach, as opposed to the all-in-one design of the TEI document. This is not really mitigated by the common practice of pipelining, or building a TEI document from separate components. The schema still specifies that it is all-in-one, and markup and internal versions are still embedded within the text. This document-centric nature of TEI contrasts with the modern data-centric world, where the focus is much more on connecting relatively smaller chunks of data (Berners-Lee 2006).

51 Interoperability also is not a goal that can be ignored simply because it is judged to be "impossible" using current technology. What matters is what users need, and they need interoperability now more than ever. Human interpretations will never be interoperable on their own, but it is possible to incorporate them into a technological structure that takes into account their variability.

52 The creation of customized SGML/XML encodings of literary and historical documents has not yet led to general sharing and collaboration in twenty-six years of trying, and the case has been made here that it never will. There is something fundamentally different about the way digital humanists encode texts that seems to make this impossible.

53 The objective of AustESE is to create digital scholarly editions that are as far as possible interoperable, and general tools to manage and visualize them. There is still much work to be done, for example, in building the stand-off editor outlined here, and tools for linking text and images need to be completed. But ideally the digital scholarly edition should be an *abstract* specification or model that can be realized in a variety of technological ways, so that it may conform to the ever-present currents of change.

BIBLIOGRAPHY

Bauman, Syd. 2011. "Interchange vs. Interoperability." In *Proceedings of Balisage: The Markup Conference 2011*. Balisage Series on Markup Technologies, vol. 7. https://balisage.net/Proceedings/vol7/html/Bauman01/BalisageVol7-Bauman01.html. doi:10.4242/BalisageVol7.Bauman01.

Berners-Lee, Tim. 2006. "Linked Data." "Design Issues: Architectural and Philosophical Points." Last modified June 18, 2009. http://www.w3.org/DesignIssues/LinkedData.html.

Blackwell, Christopher, and Neel Smith, 2012. "The CITE Architecture." *Homer Multitext Project: Documentation*. Last modified May 19, 2013. http://www.homermultitext.org/hmt-doc/cite/.

Bray, Tim, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, and François Yergeau. 2008. *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. W3C Recommendation 26 November 2008. Last modified February 7, 2013. http://www.w3.org/TR/REC-xml/.

Causer, Tim, Justin Tonra, and Valerie Wallace. 2012. "Transcription Maximized; Expense minimized? Crowdsourcing and Editing The Collected Works of Jeremy Bentham." *Literary and Linguistic Computing* 27(2): 119–37. doi:10.1093/llc/fqs004.

Crompton, Constance, and Raymond Siemens. 2012. "The Social Edition: Scholarly Editing Across Communities." Presented at Digital Humanities Hamburg, Germany, July 16–20. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/the-social-edition-scholarly-editing-across-communities/.

Cummings, James. 2007. "The Text Encoding Initiative and the Study of Literature." In *A Companion to Digital Literary Studies*, edited by Ray Siemens and Susan Schreibman, 451–76. Oxford: Blackwell.

Day, Michael. 2001. "Metadata in a Nutshell." Draft of an article published in *Information Europe* 6 (2): 11. http://www.ukoln.ac.uk/metadata/publications/nutshell/.

Dombrowski, Quinn, and Seth Denbo. 2013. "TEI and Project Bamboo." *Journal of the Text Encoding Initiative* 5. http://jtei.revues.org/787. doi:10.4000/jtei.787.

Durusau, Patrick. 2006. "Why and How to Document Your Markup Choices." In *Electronic Scholarly Editing*, edited by Lou Burnard, Katherine O'Brien O'Keeffe, and John Unsworth, 299–309. New York: MLA.

Estival, Dominique, and Nick Nicholas. 1997. "TEI Encoding and Syntactic Tagging of an Old French Text." Paper presented at the Text Encoding Initiative Tenth Anniversary User Conference, Providence, Rhode Island, November 14–16. http://cds.library.brown.edu/conferences/tei10/tei10.papers/estival.html.

Gabler, Hans Walter, Wolfhard Steppe, and Claus Melchior, eds. 1984. *Ulysses: A Critical and Synoptic Edition*. By James Joyce. New York and London: Garland.

Gabler, Hans Walter. 2010. "Theorizing the Digital Scholarly Edition." *Literature Compass* 7(2): 43–56. doi:10.1111/j.1741-4113.2009.00675.x.

Geyken, Alexander, Susanne Haaf, and Frank Wiegand. 2012. "The DTA 'base format: A TEI-Subset for the Compilation of Interoperable Corpora." Paper presented at KONVENS 2012 (LThist 2012:

First International Workshop on Language Technology for Historical Texts), Vienna, September 21. In *Empirical Methods in Natural Language Processing: Proceedings of the Conference on Natural Language Processing 2012*, edited by Jeremy Jancsary, 383–91. Vienna: Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI). http://www.oegai.at/konvens2012/proceedings/57_geyken12w/.

Goldfarb, Charles F. 1996. *The Roots of SGML—A Personal Recollection*. Accessed August 20, 2014. http://www.sgmlsource.com/history/roots.htm.

Haynes, David. 2004. *Metadata for Information Management and Retrieval*. London: Facet.

Hunter, Jane, and Anna Gerber. 2012. "Towards Annotopia—Enabling the Semantic Interoperability of Web-Based Annotations." *Future Internet* 4(3): 788–806.

Ide, Nancy, C. M. Sperberg-McQueen, Robert Amsler, Donald Walker, Susan Hockey, and Antonio Zampolli. 1988. "Proposal for Funding for an Initiative to Formulate Guidelines for the Encoding and Interchange of Machine-Readable Texts." Last modified April 1988. http://www.tei-c.org/Vault/SC/scg02.html.

Jones, Stephen E. 2014. *The Emergence of the Digital Humanities*. New York and London: Routledge.

Leibert, Marie. 2008. *Project Gutenberg (1971–2008)*. eBook. Released October 26. http://www.gutenberg.org/ebooks/27045.

Morrás, Maria. 2003. "Informática y crítica textual: realidades y deseos." In *Literatura hipertextual y teoría literaria*, edited by María José Vega, 225–41. Madrid: Mare Nostrum Communicación.

Mueller, Martin. 2011. "To Members of the TEI-C Board and Council, From Martin Mueller, chair, TEI-C Board." Accessed August 20, 2014. http://ariadne.northwestern.edu/mmueller/teiletter.pdf.

———. 2013. "TEI-Nudge or Libraries and the TEI." Guest post, *Digital Humanities blog*, Northwestern University Library Center for Scholarly Communication & Digital Curation. http://cscdc.northwestern.edu/blog/?p=872.

Nelson, Theodore Holm. 1997. "Embedded Markup Considered Harmful." October 2. http://www.xml.com/pub/a/w3j/s3.nelson.html.

Perry, David J. 2006. "Proposal to Add Ancient Roman Weights and Monetary Signs to UCS." http://www.unicode.org/L2/L2006/06173-roman-coinage.pdf.

Pierazzo, Elena. 2009. "Digital Genetic Editions: The Encoding of Time in Manuscript Transcription." In *Text Editing, Print and the Digital World*, edited by Marilyn Deegan and Kathryn Sutherland, 169–86. Farnham: Ashgate.

Porter, Dot. 2013. "Medievalists and the Scholarly Digital Edition." *Scholarly Editing* 34. http://www.scholarlyediting.org/2013/essays/essay.porter.html.

Renear, Alan, Elli Mylonas, and David Durand. 1993. "Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies." Final version, January 6. http://cds.library.brown.edu/resources/stg/monographs/ohco.html.

Renear, Alan. 2000. "The Descriptive/Procedural Distinction Is Flawed." *Markup Languages: Theory & Practice* 2(4): 411–20.

Robinson, Peter. 2010. "Electronic Editions for Everyone." In *Text and Genre in Reconstruction*, edited by Willard McCarty, 145–63. Cambridge: Open Book Publishers.

———. 2013. “Five Desiderata for Scholarly Editions in Digital Form.” Paper presented at Digital Humanities 2013, University of Nebraska–Lincoln, 16–19 July. http://dh2013.unl.edu/abstracts/ab-314.html.

Sahle, Patrick. 2013. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung*. Schriften des Instituts für Dokumentologie und Editorik, Band 9. Norderstedt: Books on Demand (BoD).

Singer, Kate. 2013. “Digital Close Reading: TEI for Teaching Poetic Vocabularies.” *Journal of Interactive Technology and Pedagogy* 3. http://jitp.commons.gc.cuny.edu/digital-close-reading-tei-for-teaching-poetic-vocabularies/

Sperberg-McQueen, C. M.. 1991. “Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts.” *Literary and Linguistic Computing* 6(1): 34–46. doi:10.1093/llc/6.1.34.

Spinazzè, Angela. 2004. “Museums and Metadata: A Shifting Paradigm.” In *Metadata in Practice*, edited by Diane I. Hillman and Elaine L. Westbrooks, 37–50. Chicago: American Library Association.

Stapelfeldt, Kirsta, and Donald Moses. 2013. “Islandora and TEI: Current and Emerging Applications/Approaches.” *Journal of the Text Encoding Initiative* 5. http://jtei.revues.org/790. doi:10.4000/jtei.790.

TEI Consortium. 1988. *Design Principles for Text Encoding Guidelines: TEI ED P1*. Last revised January 9, 1990. http://www.tei-c.org/Vault/ED/edp01.htm.

———. 2014. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.6.0. Last updated January 20 2014. N.p.: TEI Consortium. http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/.

Unicode. 2010. “Ligatures, Digraphs, Presentation Forms vs. Plain Text.” Frequently Asked Questions. Last modified September 3. http://www.unicode.org/faq/ligature_digraph.html.

Unsworth, John. 2011. “Computational Work with Very Large Text Collections: Interoperability, Sustainability, and the TEI.” *Journal of the Text Encoding Initiative* 1. http://jtei.revues.org/215. doi:10.4000/jtei.215.

Usdin, B. Tommie. 2002. “When ‘It Doesn’t Matter’ Means ‘It Matters’.” In *Proceedings of Extreme Markup Languages*. http://conferences.idealliance.org/extreme/html/2002/Usdin01/EML2002Usdin01.html.

———. 2009. “Standards Considered Harmful.” In *Proceedings of Balisage: The Markup Conference 2009* . Balisage Series on Markup Technologies, vol. 3. http://www.balisage.net/Proceedings/vol3/html/Usdin01/BalisageVol3-Usdin01.html. doi:10.4242/BalisageVol3.Usdin01.

Vulpe, Michael J. M. G., and Stephen P. Owens. 1998. *Method and system for manipulating the architecture and the content of a document separately from each other*. US Patent 5,787,449, filed June 2, 1994, and issued July 28, 1998. http://patft.uspto.gov/netacgi/nph-Parser?Sect2=HITOFF&p=1&u=/netahtml/PTO/search-bool.html&r=1&f=G&l=50&d=PALL&RefSrch=yes&Query=PN/5787449.

Zapf, Volker. 2006. *HNML: HyperNietzsche Markup Language*. http://www.hypernietzsche.org/events/sew/post/Slides%20and%20Texts_files/HNML.pdf.

NOTES

1. "Projects Using the TEI," accessed February 13, 2014 http://www.tei-c.org/Activities/Projects/.

2. Dariah, last modified July 30, 2013: http://www.daria.eu; TextGrid: http://www.textgrid.de; HuNI: http://huni.net.au; Islandora: Stapelfeldt and Moses 2013.

3. http://www.fedora-commons.org.

4. http://www.dspace.org/.

5. Last revised February 8, 2013, http://www.openannotation.org/spec/core/.

6. Digital Research Tools Wiki https://digitalresearchtools.pbworks.com/w/page/17801708/Text%20Analysis%20Tools.

7. *An Interoperable Supranational Infrastructure for Digital Editions (Interedition)*, http://w3.cost.eu/fileadmin/domain_files/ISCH/Action_IS0704/progress_report/progress_report-IS0704.pdf.

8. Scalable Vector Graphics, http://www.w3.org/Graphics/SVG/.

9. http://www.tei-c.org/Activities/Projects/

10. Base de Français Médiéval, http://bfm.ens-lyon.fr/

11. Croatian Language Repository, http://riznica.ihjj.hr/.

12. Icelandic Online Dictionary and Readings, http://uwdc.library.wisc.edu/collections/IcelOnline.

13. Medieval Unicode Font Initiative, http://www.mufi.info/.

14. http://www.markdownviewer.com/.

15. Australian Electronic Scholarly Editing, http://austese.net/.

16. http://standoffmarkup.org/.

17. Dublin Core Metadata Initiative, http://dublincore.org/documents/.

18. Metadata Object Description Schema, http://www.loc.gov/standards/mods/.

19. Encoded Archival Description Tag Library, version 2002, EAD Technical Document no. 2, http://www.loc.gov/ead/tglib/index.html.

20. http://www.neo4j.org/

21. Humanist, 2010. "inadequacies of Markup." *Humanist Discussion Group archives*, vol. 23, digest 789, 3; 795, 1. http://dhhumanist.org/Archives/Current/Humanist.vol23.txt. "piece of cake:" Wendell Piez, 30 April 2010. "five minutes work:" John Walsh, 4 May 2010.

22. The Sydney Electronic Text and Image Service, http://setis.library.usyd.edu.au/ozedits/harpur/.

23. Wendell Piez, Humanist Discussion Group Vol 23, digest 789, 3. http://dhhumanist.org/Archives/Current/Humanist.vol23.txt.

24. https://github.com/AustESE-Infrastructure/psef-tool.

25. EPUB3 Overview: Recommended Specification, October 11, 2011, http://www.idpf.org/epub/30/spec/epub30-overview.html.

26. EPUB3 Annotation. Epub-revision. http://code.google.com/p/epub-revision/wiki/ImplementationProposalAnnotations.

ABSTRACT

Recent proposals for creating digital scholarly editions (DSEs) through the crowdsourcing of transcriptions and collaborative scholarship, for the establishment of national repositories of digital humanities data, and for the referencing, sharing, and storage of DSEs, have underlined the need for greater data interoperability. The TEI Guidelines have tried to establish standards for encoding transcriptions since 1988. However, because the choice of tags is guided by human interpretation, TEI-XML encoded files are in general not interoperable. One way to fix this problem may be to break down the current all-in-one approach to encoding so that DSEs can be specified instead by a bundle of separate resources that together offer greater interoperability: plain text versions, markup, annotations, and metadata. This would facilitate not only the development of more general software for handling DSEs, but also enable existing programs that already handle these kinds of data to function more efficiently.

INDEX

Keywords: digital scholarly editions, interoperability, stand-off markup, metadata, annotation

AUTHOR

DESMOND SCHMIDT

Desmond Schmidt has degrees in classical Greek papyrology from the University of Cambridge, UK, and in Information Technology from the University of Queensland, Australia. He has worked in the software industry, in information security, on the Vienna Edition of Ludwig Wittgenstein, on Leximancer, a concept-mining tool, and on the AustESE (Australian electronic scholarly editing) project at the University of Queensland. He is currently a Research Scientist at the Institute for Future Environments, Queensland University of Technology.