DHQ: Digital Humanities Quarterly
2017
Volume 11 Number 2

Some principles for making collaborative scholarly editions in digital form

Peter Robinson <peter_dot_robinson_at_usask_dot_ca>, University of Saskatchewan

Abstract

"Textual Communities" is a new system for managing and performing all aspects of an online collaborative scholarly editing project. It permits mounting of document images and offers page-by-page transcription and display, with the facility for project leaders to recruit and manage transcribers and other contributors, allocating and reviewing transcription work as it is done. Most distinctively, Textual Communities is built on a comprehensive model of scholarly editing, enabling both "document" (page-by-page) and "work" (intellectual structure, or "entity") views of the texts edited. Accordingly, multiple texts of a single work, or part of a work (an entity) may be extracted and compared, using an embedded installation of CollateX. While completely conformant with Text Encoding Initiative guidelines, Textual Communities goes beyond TEI and XML in its ability to handle multiple overlapping hierarchies within texts. This paper will outline the thinking behind the development of Textual Communities, and show examples of its use by several major projects.

Introduction: scholarly editions in the digital age

While one may dispute how "revolutionary" scholarly editions in digital form may be as compared to their print counterparts [Robinson forthcoming], we can agree that the onset of digital methods has considerably broadened the ways in which editions may be made and distributed. We[1]can now contemplate editions made by wide collaborations of editors, transcribers, indexers, commentators and annotators, all working across the internet. We are now accustomed to seeing editions providing multiple interfaces, with distribution ranging from strictly-controlled paid-for access to access open to anyone with an internet connection. 1

These possibilities require us to ask new questions. In terms of the edition as product, as something made: how can we most usefully characterize the fundamental intellectual components of a scholarly edition in the digital landscape? In this article, I consider this question in terms familiar to scholarly editors: the concepts of document, text and work. In terms of the edition as a process, as something we make and use: who are "we"? how may we relate to each other, both as creators and as readers? Indeed, digital methods open up a yet more radical possibility: that readers may become creators, as the edition becomes an ever-open field of interaction, with readers contributing to its continuing remaking. In what follows, I express these two perspectives as "axes". 2

This is not simply a matter of describing editions and how they may be made. We have choices, more than ever, in the digital age, as to where we locate our edition against these axes. An edition can be severely limited, or richly substantive, both in terms of what it contains and who may use it and how they may use it. In the latter part of the article, I describe an editorial environment, "Textual Communities", designed with these axes in mind. It goes without saying that Textual Communities, like the editions it might make, is subject to never-ending transmutation. 3

The axes of scholarly editions

Every scholarly edition may be placed on two axes, representing (as it were) the longitude and latitude of editing. The first axis is along the familiar continuum of document/text/work: is the edition devoted to a particular document (as for a modern genetic edition of an authorial manuscript)? Or is it oriented towards presentation of a work found in many 4

documents (as for an edition of the Greek New Testament, or a medieval work found in many manuscripts)? The second axis is along the range of relationships between the editor and the editions' audience. Is the edition made by a specialist scholar and intended for a narrow and specialist audience, not to be read but to be used as a resource for further scholarly work? Or is it made by a non-specialist and intended for the broadest possible audience, to be read rather than studied? Does its design and implementation permit its endless re-use, so that readers may in turn become editors, or other editors may take it and reforge it for their own purposes and audiences? Note too that as in geographical coordinates, the place of an edition on one axis is independent of its place on the other axis. An edition intended for the general reader may be based on all the documents, or on just one; an edition intended for a specialist reader may also be based on all the documents, or on just one. An edition based on a single document, or on many, might be designed to permit other editors to take and repurpose what is made in ways unforeseen by its original makers.

Figure 1. The document/work/text axis

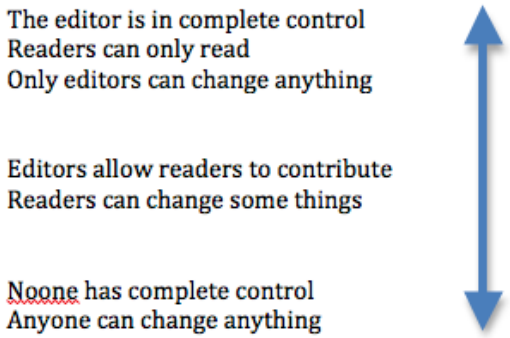

Figure 2. The editor/audience axis

5 There is no novelty about these two axes. Editions made long before the advent of digital methods may be referenced along these axes. However, digital methods have widened the range of choices along each axis, and also altered the balance between these choices, as they favour one choice over another. For example: digital imaging has made it possible to make full-colour facsimiles for a very low cost (as little as a few cents a page) and to distribute them over the internet to any one with an internet connection and a computer, at no cost to the reader. The emergence of successful methods for encoding a document page-by-page combined with the availability of high-quality digital images has favoured the making of a particular kind of edition, oriented towards the document rather than the work. Along the other axis: the editorial decision about the audience of the edition has been complicated by the emergence of funding agencies as the primary sponsors of editions, and by the emergence of centres substantially funded by these same agencies as the primary place of making editions. In these circumstances, the primary motivation of an edition is not to reach an audience, whether narrow or broad, but to satisfy the funder. This has also led to the emergence of specialists in computer methods as leaders in the making of editions, rather than specialists in (for example) the texts themselves.

6 One of the most influential commentators on scholarly editing in the digital age, Jerome McGann, has frequently cited William Morris's observation that (in McGann's wording) "you can't have art without resistance in the material."[2] There is danger in failing to resist: in editing, as in art, to do what is easiest may not be to do what is best. Following the scenario sketched in the last paragraph, we have seen the making of numerous "digital documentary editions" [Pierazzo

2011], typically created with considerable funding and support from a digital humanities centre [Sutherland 2010] [Maryland 2014]. The coincidence of cheap digital images, powerful encoding, and significant funding has resulted in several remarkable editions. However, there is a danger of imbalance, if many editions are made in a narrow band along the document/text/work axis, and for an ill-defined audience. Resistance in the materials entails not doing what is easiest. In scholarly editing terms: resistance means deciding where on the document/text/work continuum your edition should be located, not simply placing it where the technology makes it easiest, and identifying what audience you wish to serve, not simply satisfying your funder and your own inclinations.

Documents, texts and communicative acts

7 Accordingly, the first task of any editor is to know what is meant by the terms document, text and work, and how the edition may be located with reference to these terms. Over the last decades, several scholars (notably Peter Shillingsburg, Paul Eggert and Hans Gabler; see footnote 1) have debated the valency of these terms, and I have summarized their arguments and presented my own definitions in two articles [Robinson 2013a], [Robinson 2013b]. In summary: a *document* is the material object upon which marks are inscribed: a manuscript, a book, a clay shard. A *text* is the communicative linguistic act which the reader deduces as present in the document, represented by these marks. A *work* is a set of texts which are hypothesized as related, in terms of the communicative acts which they present.[3] Thus: the Hengwrt manuscript is a document preserved in the National Library of Wales, comprising some 248 folios bound into 32 quires. This manuscript contains a text which we recognize as an instance of Geoffrey Chaucer's *Canterbury Tales*. Further, we know another eighty three manuscripts and four print editions dating from before 1500 containing texts of the *Tales*, and (of course) many editions, adaptations and translations dating from after 1500. We speak of the "work" as all these together. In our common thought, we conceive the work as something beyond all these physical documents: as the creative act conceived and executed by Geoffrey Chaucer in the last decades of the fourteenth century.

8 These definitions have many implications. First: there is a clear division between document and text. The text is not simply the marks in the document. It is the communicative act that I, the reader, identify as represented by those marks. The difference may seem slight. It is critical. When we record the text of the document, we are not simply saying: that mark is an "i", this next mark is a "t"; we have "it". We see first a set of potentially meaningful marks in the document. We hypothesize: these marks may represent a communicative act. [4] They are not (say) marks left by grubs crawling under the bark of a tree [Eggert 2010]. Someone made these marks to communicate something. We then identify the language and writing system to which these marks belong. We identify these marks as letters, composing the word "it". As we examine the marks, the communicative act takes shape in our minds: the words resolve into sentences, into verse lines, paragraphs, as part of something we know as the General Prologue of the *Canterbury Tales*. This communicative act has a double aspect. One aspect is the disposition of the marks in the document: exactly where on the page they appear; the combination of strokes which compose each letter. The other aspect is that of the components of the communicative act: for prose, as constituted by a sequence of sentences within paragraphs; for verse, as lines within stanzas. It is not simply a sequence of words: it is structured, capable of division and labeling.

9 Normally, these processes of recognition happen so quickly, so instinctively, that they do not appear like thought at all. We see a text on a page and we read it. We think the marks on the page are the text, and hence that recording the text is a mechanical act of transposing those marks from one medium (the page) to another (now, usually, an electronic file in a computer). Accordingly one sees statements implying (or indeed asserting) that a transcription can somehow not be "interpretive", and hence aspire to some kind of "objective" status. As an example: in a discussion on the Text Encoding Initiative list in February 2014, several participants routinely described transcription of the communicative act, in terms of paragraphs, sentences, identification of names and places within the text, as "interpretive" (or "interpretative"), while recording exactly where the text appears on the document page was described as "non-interpretive". The Text Encoding Initiative even has distinct elements for the two types of transcription: "interpretive" transcripts are held in `<text>` elements; "non-interpretive" transcripts are held in `<sourceDoc>` elements.[5] This distinction led Barbara Bordalejo to ask tartly in the course of that discussion: "are you suggesting there are transcriptions that are not interpretive? How do you distinguish them?"

Encoding text as document and as communicative act

10 In the definition of text here offered, all is interpretation: there is no such thing as a “non-interpretive” transcript. Further, this definition stipulates that text has two aspects: it is both marks upon paper (corresponding to TEI <sourceDoc>) and it is the components of a communicative act (corresponding to TEI <text>). Both aspects may be expressed as hierarchies. The document hierarchy consists of the book, divided into quires, divided into pages, divided into writing spaces: columns, lines, margins. The components of the communicative act may also be expressed as chapters divided into paragraphs divided into sentences, or poems divided into stanzas divided into lines. The two hierarchies are completely independent of each other. The General Prologue may be written across several quires, or contained in only one; it may spread across as few as six folios, or as many as sixteen. Of course, the hierarchies overlap. Paragraphs continue across page breaks, lines of verse across line breaks. In the world of documents, this is no problem. The text of the communicative act runs across quires, pages, line breaks in an orderly and straight-forward manner. We are so used to this that we do not notice it. We skip from page to page, across footnotes, past catchwords, page numbers, running heads. Sometimes, the two hierarchies coincide. The book opens a new story, a new chapter opens on a new page, a new section a new volume, before once more the hierarchies diverge, and each runs their separate course, to the end of the book and the end of the story. The codex, whether in scroll, manuscript or print form, is superbly fitted to carry the text of communicative acts. A New Testament gospel might fit neatly in a single small codex; the whole New Testament in a larger one, or split across several codices. This overlapping, this sliding of one hierarchy across another, this disposition of this printing across many volumes in one instance, or in just one in another instance, is so common that an editor might note it briefly, and move on.

11 But while it is straightforward to represent a communicative act in a document, it is not at all straightforward to represent both the components of a communicative act and of a document in a single electronic representation – and particularly not in a single electronic representation which conforms to the norms of the Text Encoding Initiative, the gold standard of encoding for humanities texts. The XML (“eXtensible Markup Language”) specification requires that content objects within an XML document conform to a single hierarchy. Accordingly, it is a simple matter to represent either the document hierarchy (books, quires, pages, lines) or the communicative act components hierarchy (poem, stanzas, lines; story, chapters, paragraphs). But it is not at all simple to represent both hierarchies in a single XML document. [6] Over the twenty years of encoding of texts of primary sources using the TEI guidelines, scholars have used various devices to circumvent this problem. In the “P3” version of the guidelines, the chapter on encoding of primary sources suggests that one should represent the communicative act component hierarchy exactly and fully, by identifying each part of the communicative act (each paragraph, each verse line) with a discrete segment of the TEI document (thus, a <p> or <l> element), and then nesting the segments within other segments, so that <p> elements are contained within elements, just as paragraphs are contained within chapters.[7] XML (like its predecessor, SGML) is optimized for representing a single “ordered hierarchy of content objects”: but it does also have a means of recording other information about the encoded text, in the form of “empty elements”, otherwise known as “milestones”. Accordingly, in a TEI document one might record the communicative act hierarchy as the primary hierarchy, and then represent the document hierarchy as a sequence of milestone elements: <pb/> and <lb/> elements for pages and lines. In this ordering, the <pb/> and <lb/> elements, unlike the and <p> elements, hold no content: they state where page breaks and line ends are relative to the text of the communicative act within the document. The result is that the components of the communicative act are represented completely, and one may readily use all the tools available in the XML community, optimized for dealing with hierarchies of content objects, to manipulate the document. However, the material document which holds the text is represented far less adequately. One might record the larger features of the document – the number of pages within it, the number of lines within each page – and record too exactly the page and line breaks that occur within the text. But it will be difficult to represent more complex phenomena, such as a single page which contains multiple writing spaces, each containing text in a complex relation with texts in other writing spaces. Further processing of the document according to this second hierarchy is complex, and often impractical.

12 The P3 fundamental logic – that one identifies the components of the communicative act (sometimes referred to as “intellectual” or “logical” structure) as the primary hierarchy of its XML representation, and record document features as milestone elements – is used in countless TEI-based encodings of primary textual materials, including several scholarly

editions in digital form (such as those made by myself or in which I was involved, e.g. [Robinson 2004] and [Shaw 2010]. This logic prioritizes representation of the components of the communicative act over the physical document, and so is well-suited to situations where the disposition of the text in the document is either straight-forward, as in many medieval manuscripts or printed books, so that it may be adequately captured through sequences of page and line-breaks alone, or is perceived as relatively unimportant. However, there is an important class of documents where the disposition of the text on the page is both complex and significant. This applies particularly to authorial manuscripts, where authorial revision is expressed through multiple acts of writing within a page, from which editors must construct a text or texts by decryption of the sequence of revisions embedded in these multiple writings. In these cases, the P3 system is inadequate. Further, continuing from the last decades of the last century, scholarly editors have become increasingly interested in the "material text", following the ground-breaking writings of Donald McKenzie [McKenzie 1999] and Jerome McGann #mcGann1983, and continuing through many others.[8] Thus, it became increasingly important to many scholars to represent as exactly as possible the document page and the text upon it, with a fullness and precision which the P3 system could not achieve. In response to this need, the TEI convened a working group to prepare encodings for the making of documents where representation of the document was paramount. This resulted in a new Section 11.2.2, first issued in "version 2.0.0" of the P5 Guidelines in December 2011. This section introduced a new high-level <sourceDoc> element, specifically to carry the "embedded transcription" described in this section. This "embedded transcription" is described as "one in which words and other written traces are encoded as subcomponents of elements representing the physical surfaces carrying them rather than independently of them". The examples and the accompanying documentation make very clear exactly what is meant by this: that the marks upon the page are interpreted as words completely independent of any sense of their being part of a communicative act. Thus, the letters and words of the page are placed within the page hierarchy, in a series of elements which may be nested within one another: the page as <surface>, which might contain a <zonee> (a column, a writing area), itself containing <line> and <sege> elements, which might contain the words themselves. There is no place here at all for recording information about the text as a structured communicative act. Instead, the Guidelines suggest that information about the text as communicative act should be recorded in a separate <texte> element, parallel to the <sourceDoc> elements. In theory, this is a better solution than the rather makeshift procedure adopted by P3. In practice, it is extremely difficult to maintain two distinct transcriptions, and to maintain the complex sets of links between the two.[9]

13

The Shelley-Godwin archive shows the power of document-based encoding [Maryland 2014]. A feature of this new encoding is that it provides for explicit statement of the revisions within each document page and their sequence. One can see (for example) exactly what Mary Shelley wrote, and what Percy Shelley wrote. One can read the transcription in parallel to each page facsimile, with each element of the transcription mirrored in transcript and page: a considerable technical feat. However, what is excellent for these materials – a classic instance of "genetic texts", through which one may see the authors (in this case) forging the text a phrase at a time – may not be appropriate for other editions. While the TEI guidelines recommend that a parallel encoding of the text-as-structured-communicative-act be made alongside the encoding of the text-as-document, in practice editions may not follow this advice, and indeed the Shelley-Godwin archive does not do this.[10] The result is that while one can see precisely the changes within each page, the failure to encode the components of the communicative act within each document makes it extremely difficult to see the changes between one document and another. Indeed, the rigid segmentation of the document into pages makes it impossible to record a change which spans across a page boundary. For example: one finds on fol 5v of Fair Copy Notebook C1. c. 58 and on fol 73r of Draft Notebook B. c. 57 versions of the end of chapter 22. But to locate these passages one is reduced to using the search engine to discover parallel texts: not a very efficient procedure.

Representation of both work and document: the DET system

14

This defect brings us to the third element of the document/text/work triumvirate: the work. The definition of work given above – that a work is a set of texts which are hypothesized as organically related, in terms of the communicative acts which they present – depends on identification of an instance of the communicative act and its components in any one document (e.g. this is the General Prologue in the Hengwrt manuscript) and then identification of other instances of related communicative acts in other documents (this is the General Prologue in the Ellesmere manuscript, in the Caxton printings, in the Riverside Chaucer). Because we identify the components of the communicative act in any one

document, we can compare its instantiation in that document with its instantiation in any other. If we reduce our notion of text to simply words in documents, we have no means of asserting relations between documents apart from the happenstance of some words recurring in different documents (as I was able to use the Shelley/Godwin archive search engine to discover that folios in different notebooks had similar words to the end of Chapter 22 in various print editions). This is unsatisfactory, to put it mildly. It provides no means of linking translations, or radical rewritings. We can assert (to give an extreme example) that the many hundred manuscripts of the medieval *Visio Pauli* are related, through many recensions in many languages, many of which have not a single word in common, because they share structure, subject, theme, motifs and details, and because we can trace the historical growth of the tradition across time and space and from document to document.[11] I can assert that both the Sion and the Merthyr manuscripts contain versions of the *Canterbury Tales*, even though there is not one line of the Tales in common to both (Sion holds the sequence Clerk's Tale-Summoner's Tale, Merthyr has part of the Nun's Priest's Tale and link), as surely as Darwin can assert that two multi-segmented organisms are both barnacles, even though they have not a single segment in common. Darwin can assert this by showing that both organisms descend from an ancestor which had both sets of segments. I can show that both manuscripts descend from other manuscripts which had both sets of tales.

According to these definitions, then, a fundamental requirement of scholarly editing is that both aspects of a text are recognized: both the text as marks upon a document, and the text as structured communicative act. To put it at its simplest: we need to be able to say that the words "Whan that Auerill with his shoures sote" are found in a particular space on the first folio of the Hengwrt manuscript, Peniarth 392D in the National Library of Wales, and that these words are also the first line of the General Prologue of the works we know as Geoffrey Chaucer's *Canterbury Tales*. Over the last years, with the help of many people, I have been developing a formal system for describing documents, texts and works to enable just this. In essence: we need a scheme for labeling all three elements, that will allow us to identify every part of each element, in every document, text and work, and assert too how they relate to each other. [12] 15

We call this scheme "documents, entities and texts" (DET), using the term "entity" to refer to the unique labels we give each component of a communicative act. [13] The labeling system we employ is based on the Kahn/Wilensky architecture [Kahn 2006]. Like Kahn/Wilensky, we use the familiar uniform resource name ("urn") notation to hold each label. Following Kahn/Wilensky, we separate the label into two parts: a naming authority, and the name given by that naming authority to the object. Thus, in Kahn/Wilensky the handle "berkeley.cs/csd-93-712" gives the naming authority as "berkeley.cs", and "csd-93-712" is the name given by that authority to a particular object. In full urn form, this is expressed as 16

```
<URN:ASCII:ELIB-v.2.0:berkeley.cs/csd-93-712e>
```

In our system, we adopt the use of "/" to separate the naming authority from the name, and we further specify that the name must be composed of at least one of the key words 'entity' and 'document' and of one or more key value pairs, separated by the ":" delimiter.

Applying this to a document: 17

`TC:USask:CTP2/document=Hengwrt` - indicates that the naming authority TC:USask:CTP2 has given this document the name "Hengwrt" 18

`…/document=Hengwrt:Folio=1r` – indicates Folio 1r of the Hengwrt manuscript 19

`…/document=Hengwrt:Folio=1r:line=2` – indicates line 2 of Folio 1r of the Hengwrt manuscript 20

Applying this to an entity, that is to a named component of a communicative act: 21

`TC:USask:CTP2/entity=Canterbury Tales` – indicates that the naming authority TC:USask:CTP2 has given this entity the name "Canterbury Tales" 22

…/entity=Canterbury Tales:Section=General Prologue – the General Prologue of the Canterbury Tales

…/entity=Canterbury Tales:Section=General Prologue:line=1 – line 1 of the General Prologue of the Canterbury Tales

In full urn notation, the document would be

<urn:DET:TC:USask:CTP2/document=Hengwrt>; the entity would be <urn:DET:TC:USask:CTP2/entity=Canterbury Tales>

We have defined a “text” as a communicative act, which comprises both document (the material upon which it is inscribed) and entity (the components into which it might be divided). Accordingly, a text of any one communicative act in any one document is the collocation of the entities and of the document for that text. Thus, for the text of the Canterbury Tales in the Hengwrt manuscript:

```
TC:USask:CTP2/document=Hengwrt:entity=Canterbury Tales
```

For the text of the first line of the General Prologue on the second line of folio 1r of the Hengwrt Manuscript:

```
…/document=Hengwrt:Folio=1r:line=2:entity=Canterbury Tales: Section=General
Prologue:line=1
```

The power of this system should be immediately apparent. We can, from this naming alone, identify all manuscripts which contain the *Canterbury Tales*; all manuscripts which contain the General Prologue; all manuscripts which contain the first line of the General Prologue. Or, in reverse: we can say, for any one manuscript, exactly what parts of the *Canterbury Tales* it contains; we can say, for any page in any manuscript, what lines of what part of the Tales it contains; we can say, for any line or space in any page in any manuscript exactly what words of what part of the Tales it contains. Note that the document and entity naming is completely hierarchical: each successive object within the sequence of name/value pairs must be contained within the preceding object. Line one is part of the General Prologue, which is part of the *Canterbury Tales*; the second line is on Folio 1r which is part of the Hengwrt manuscript. Note too that the system can cope with prose and other texts where communicative acts span across lines and pages. Paragraph 162 of the Parson’s Tale in the Corpus Christi 198 manuscript of the *Tales* begins on line 36 of folio 272r, and continues on the first two lines of folio 272v. This can be represented as follows:

```
…/document=Corpus:Folio=272r:line=36:entity=Canterbury Tales:
Section=Parson’s Tale:Segment=162
```

```
…/document=Corpus:Folio=272v:line=1:entity=Canterbury Tales:
Section=Parson’s Tale:Segment=162
```

```
…/document=Corpus:Folio=272v:line=2:entity=Canterbury Tales:
Section=Parson's Tale:Segment=162
```

Implementing DET: Textual Communities

30 Theory is one thing; implementation is another. The basic outline of this scheme was prepared by myself, with advice and help from Federico Meschini and Zeth Green, in 2008-2009, and presented first by myself and Green at a symposium on Collaborative Scholarly Editing in Birmingham in 2009, and then by Meschini and myself in a paper presented to the ADHO conference in London in 2010. Following suggestions at a meeting of the InterEdition project in Pisa in 2009, we first experimented with expressing this scheme in the form of an ontology. This was successful, as proof of concept: we could indeed connect documents, entities and texts via RDF classes and properties. [14] This helped persuade us that the concept was fundamentally sound. However, implementation of even a basic working prototype would take considerable effort and resources. In late 2010 I moved from the University of Birmingham, UK, to the University of Saskatchewan, Canada, and a considerable motive was the prospect of adequate funding to create a real editing system, based on these concepts. Such a system was needed also to support my own editorial work, particularly on Geoffrey Chaucer's *Canterbury Tales*.

31 With funding initially from the University of Saskatchewan (2010-2011), then from the Canada Foundation for Innovation (2011-2014) and now from the Canadian Social Sciences and Humanities Research Council (2014-), we have made a collaborative editing environment, "Textual Communities", built on the documents, entities and texts definitions here explained. A full technical description of the components of the Textual Communities environment is beyond the scope of this article. In brief:

- Although the DET system is designed to support full hierarchies for both document and communicative act, and has no difficulties with overlapping hierarchies, we based the system on TEI encodings which cannot support overlapping hierarchies. Partly, this was because we had many thousands of pages of transcription already in TEI encoding. Also, we knew from years of experience that use of the P3 primary text model, encoding the communicative act hierarchy as the main hierarchy and recording the document hierarchy through milestones, could yield useful results.
- Our first intent was to work through RDF, and so create an RDF repository of materials accessible via SPARQL and other RDF tools. Very quickly, we realized that the RDF tools then available could not support our aim of a large real-time editing environment. (This may change as new RDF tools are developed.) The tools were immature and did not scale well, and we had significant performance issues even with small amounts of text. Hence, we moved to use of a relational database for back-end storage of all data. Decades of development have made relational databases robust, responsive and fast, with a multitude of tools for management and for web server interfaces. We are currently moving from a relational database to a MongoDB system. JSON (Javascript Object Notation) has become our central internal representation of data, and the optimization of MongoDB for JSON objects maps well to our data.[15]
- The core of our implementation of the DET scheme with a database backend is the use of the TEI `<refsDcl>` element to map any TEI document to a DET scheme. Here is a fragment from a refsDcl declaration for the Canterbury Tales project:

```
                                <cRefPattern
matchPattern="urn:det:TC:USask:CTP2/entity=(.+)"
replacementPattern="#xpath(//body/div[@n='$1'])"></cRefPattern>
```

- Here, the “replacementPattern” attribute declares that every top level element within the document is to be mapped to an entity. The entity will be given the name of the ‘n’ attribute, thus is associated with the entity name “General Prologue”. The matchPattern attribute declares exactly what entity this ‘n’ attribute will be associated with: here, the entity itself. Taken together, the system now understands that when it sees as a top level , it associates that and its contents with the entity “urn:det:TC:USask:CTP2/entity=Canterbury Tales”. In essence: these refsDcl expressions are used to slice the whole document into entity and document chunks, to associate each chunk with a document and entity name, and hence each chunk of text in the document is linked to its document and entity and this information stored in the database.[16]

32 Through this system, we have now been able to store some 40,000 pages of manuscript transcription and images in the Textual Communities implementation at the University of Saskatchewan: see www.textualcommunities.usask.ca (particularly, for the Canterbury Tales project, see http://www.textualcommunities.usask.ca/web/canterbury-tales/viewer). This is now being used routinely by editors and transcribers in six substantial projects.[17] We are currently testing and refining the system before full public launch.

33 In its current form, Textual Communities does not (and cannot) go as far as we want towards a full representation of both aspects of the text, as document and as communicative act. This limitation arises from our use of the TEI as the base form for text representation. While the structure of each communicative act can be fully represented in XML, and hence in Textual Communities, the ability of a single XML document to represent only one primary hierarchy means that because our documents in Textual Communities choose to make the structure of the communicative act the primary hierarchy, then we are limited in our representation of the document hierarchy. We use (as do most TEI projects representing documents) sequences of the omnipresent <pb/>, <cb/> and <lb/> elements to represent the document hierarchy, and for the great majority of our documents and our purposes, this gives satisfactory results.[18] Note that the limitation is not in the DET scheme, and indeed the current move of Textual Communities to a JSON-based architecture will also remove this limitation within the system “backend”.The problem then will lie only in the XML structures we are currently using in the editorial and display interfaces. However, despite this, Textual Communities goes further than any other online collaborative editing environment known to me in its support for both the document and structured communicative act aspects of the text. Everyone of the more than thirty editing systems listed by Ben Brumfield at http://tinyurl.com/TranscriptionToolGDoc is either very limited in its support for recording communicative act components, or supports page-based transcription only.[19]

The second axis of scholarly editions: editors and readers

34 The definitions of document, text and work here offered, as the foundation of the DET system, are valid for any text of any period, and might have been offered by any scholar at any period. Documents vary in form, from inscription on stone to computer file, but these definitions and these relations hold, no matter what the medium. The digital age and the stringent mandates of computing systems require that they be defined more precisely than before: but if the concepts are valid, they were valid long before the invention of the computer. As earlier remarked, the advent of digital methods has favoured the making of some kinds of edition over others, but the fundamentals have not changed. However, this is not the case for the second axis of scholarly editions: the range of relationships between the makers of editions and their audiences. This has not just altered the balance between one kind of edition and another. It has created many more kinds of relationship between editor, edition and audience, enabling radically new kinds of edition.

35 It is now fundamental to the web that every reader may be a writer. The rise of social media means that communication across the web happens in every direction: now, routinely, every newspaper article on the web comes with a comments section, often more interesting than the article itself (though, mostly, not). The rise of “crowd-sourcing” leverages individual activity into collective movements: crowd-sourcing has produced remarkable results in areas as diverse as investigating the expense claims of politicians to transcription of museum labels. There have been several ventures into the use of crowd-sourcing for editorial purposes, notably the Transcribe Bentham and Easter 1916 projects [Causer

2012]; [Trinity College Dublin 2014] . Crowdsourcing raises the possibility of editions which are not made by a single editor, or even a group of editors, but by many people who may have no formal relationship with each other, and indeed nothing in common except a shared interest in a particular text. Further along the editor/audience axis, the ways in which an edition may be distributed to its readers and used by them are also vastly changed. While many readers may want just to read a text, others may wish to take the text of a document, add information to it, alter it, enrich it, correct it, combine it with other texts, and then republish it. Others in turn might take up this republished text and alter it still further, in a never-ending chain.

36

It has to be said that the scholarly editing community, up to now, has been very slow in responding to these new potentials. With respect to crowdsourcing: the Bentham and Easter 1916 enterprises are not “crowdsourced” editions. Rather, the framework of each edition, the flow of work, and all significant decisions concerning transcription systems and the distribution of the product of the editions are made by a small group of academic editors, as has always been the case for scholarly editions. Any reader is invited to contribute transcriptions, and the Easter 1916 system allows readers to go further, and contribute their own documents. But the reader’s role is strictly limited, and the Bentham project even prevents the reader changing a transcript he or she has made after it has been “locked” by an editorial supervisor. T-Pen and other systems do offer much more freedom to the editor, but at the cost of a very limited encoding of the structure of communicative acts. Further, almost all scholarly editing digital projects severely restrict how their output might be used. The Jane Austen Fiction Manuscripts project [Sutherland 2010] has a whole page bristling with restrictions: the site and everything on it is protected by copyright, no derivative works are allowed, the editor asserts her “moral right to be recognized as author and editor of aspects of this work” (it is not explained what “aspects” mean), “individual, non-commercial” use is permitted, but “All other use is prohibited without the express written consent of the editor. Any requests to use the work in a way not covered by this permission should be directed to the editor.” Indeed, most projects, while not going so far as the Jane Austen project, do invoke the “non-commercial” clause of the Creative Commons license. The effect of the “non-commercial” restriction, as many have observed [Möller 2005], is not just to restrict the republication of online materials by commercial publishers: it is actually to make it nearly impossible for anyone, commercial or non-commercial, individual or corporate, to republish those materials on the web. The problem is the ambiguity of what is “commercial”, what is “non-commercial”, in the web. If you publish your edition in a university website, which also sells university services, then it might be deemed commercial. If you publish it in your own website and this happens to provide links to other sites which sell anything, or even just belong to a commercial entity, then it might be deemed commercial – even if you and your edition did not create those links. For these reasons, many people will not touch any materials covered by the “non-commercial” license, in any circumstances. Indeed, among the many online digital materials which have been created by humanities scholars very few are both free of the non-commercial restriction, and actually made readily available for free re-use and re-publication.

Editions by everyone, for everyone

37

Full realization of the possibilities of readerly involvement in the making and use of editions depends on the materials being available for re-use and re-publication without restriction. This applies at both ends of the edition spectrum. One cannot reasonably expect that people who contribute transcriptions and other materials to an edition will be willing to do so if they cannot make use of their own transcriptions — or indeed, if they see that the editors are limiting the ways the transcriber’s work may be used, for the benefit of the editors. And of course, if you cannot be sure that you can freely distribute your own work on materials derived from an edition, then you will likely look elsewhere. For this reason, Textual Communities mandates that any materials created on the site must be made available under the Creative Commons Share-Alike Attribution licence (CC SA-BY); all software created by the project is also available as open source at https://github.com/DigitalResearchCentre.[20] The “attribution” requirement mandates that anyone who worked on the creation of the materials must be acknowledged, at every point along the publication chain. The “share-alike” requirement mandates that the materials, no matter how altered, must be made available under the same terms as they came to the user. This does not prevent a commercial publisher taking the materials, altering them, and then making them available as part of a paid-for publication: just that somehow, the publisher must make those altered materials available, for free (for example, by deposit on a public webserver). Further, Textual Communities provides an open API (Application Programmer’s Interface) that makes it possibly for anyone with reasonable computer skills to extract any

texts from editions within the Textual Communities system in just a few lines of code (see http://www.textualcommunities.usask.ca/web/textual-community/wiki/-/wiki/Main/The+API+Basics).

38

However, requiring that materials be made freely available is pointless if those materials cannot be made in the first place. Some two decades into the digital revolution, and people routinely send emails, create Word documents, spreadsheets and Facebook pages. But paradoxically, it is no easier (and indeed, arguably much harder) to create a scholarly edition than it was two decades ago. It is certainly harder if one takes the TEI advice, to create two transcriptions corresponding to the two aspects of text as document and text as structured communicative act and link them together. Even in cases where one is not going to make two parallel transcripts: the renowned complexity of the TEI guidelines, and the webs of software and hardware needed to turn a TEI document into an online publication continue to require that anyone who wants to make a scholarly edition in digital form must both master a range of special skills and have access to considerable technical resources. The effect has been to limit drastically the number of people who can make scholarly editions in digital form: in effect, to relatively few people typically at a few digital humanities centres. Hence, a key aim of Textual Communities is to make it possible for scholars to make digital editions without having to master more of the TEI than is necessary for their particular task, and with no need for specialized technical support. Further (perhaps over-ambitiously) we would like it to be possible for an edition made with this system to be placed anywhere along the range of relationships between the makers of editions and their readers. This means that as well as make it as easy as we can to use, we need to support all this range. Thus, in Textual Communities one can create a community where everyone is an editor, everyone can freely change what everyone else does, and everyone can take whatever is done and use it in any way they wish. Or, an editor can allow only the people he or she invites to collaborate in making the edition, and can insist that every page published on the edition must be approved by an editor before publication.

39

In order to support this range of roles, and to encourage community building and partnerships, Textual Communities is based on social media software, specifically on the LifeRay implementation of the Open Social software suite, itself used by Google as a foundation of its “Google Plus” social network.[21] This allows every community within Textual Communities to have its own Wiki, Blog, Bulletin Board and Chat facility. The screenshot below shows how Textual Communities appears to a user within the Canterbury Tales community:

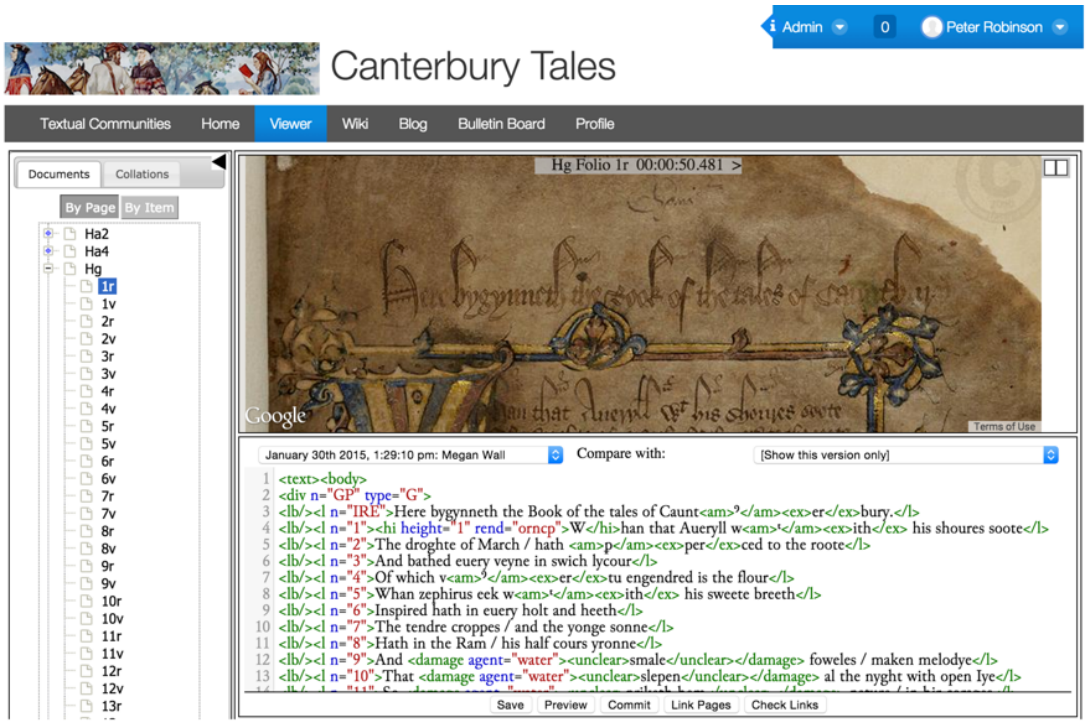

Figure 3. The Canterbury Tales community within Textual Communities: access contents by page

At the top of the screen, the “Wiki”, “Blog” and “Bulletin Board” links take the reader to the wiki, etc., for this community. To the right, we see an image above of the manuscript page; below that, the transcription of this page, in the last version saved by the transcriber. Notice that the “Compare with” tool allows the transcriber to compare different revisions of the transcription. This system does not attempt to hide the XML: we think it helpful for the editor and transcribers to see exactly what encoding is being applied to the document. Nor have we had any difficulty with transcribers at any level, including undergraduates, understanding and using the XML we use in these page transcripts (observe that, as a fully-compliant TEI implementation, any valid XML may occur within the transcripts). Note too the use of explicit `<lb/>` elements at the beginning of every manuscript line to structure the document. The buttons at the base permit the transcriber to preview the document, showing it without the XML and formatted for ease of reading, to save the transcript, and carry out various other editorial activities (including, “Link Pages”, which allows the editor to connect text which flows across the page boundaries). 40

At the left of the screen, you can see the table of contents, showing the document page by page. This table of contents is generated directly by Textual Communities from the XML, following the schema for document elements set out in the `<refsDcl>` element. If you click on the “By Item” tab, the table of contents changes: 41

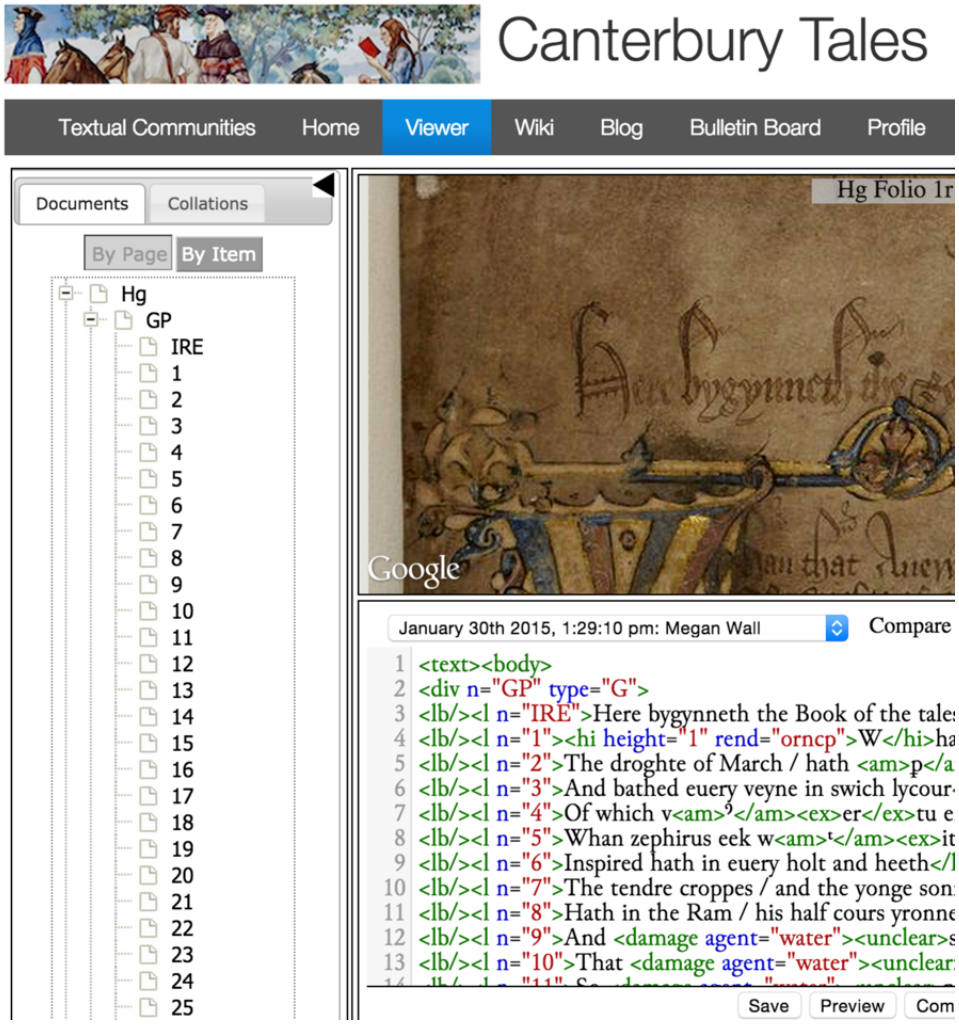

Figure 4. The Canterbury Tales community within Textual Communities: access contents by entity

Now, for the document Hg, we can access its contents by entity: that is, by the components of the communicative act. Thus, it first contains the General Prologue, which itself contains a sequence of line entities: first the initial rubric (“IRE”) and then the first and following lines. Again, the information about the entities is generated directly by Textual Communities from the XML, following the schema set out for the textual entities in the `<refsDcl>` element. Finally, the “Collations” interface allows the reader to see the collation of the text in all the documents. 42

Figure 5. The collation of the eighth line of the General Prologue

The collation here is built on the CollateX system, here extended by the addition of regularization and other facilities for adjusting the collation. 43

Conclusion

Textual Communities, like any computing system, is in constant development. We have not yet announced it publically, and will not do so until we are fully satisfied as to its robustness and usability. We are also aware that the demands of comparatively few users (six major communities, some one hundred and fifty active transcribers and editors) already place considerable strain on the current installation, on a virtual server at the University of Saskatchewan. We are both translating the system to the MongoDB backend and moving it to the Compute Canada cloud (one of the first Digital Humanities projects to be hosted on this service, hitherto devoted to “hard science” data). However, it is not at all our aim that the whole world, or at even a significant part of the whole scholarly editing community, should use Textual Communities. We are more concerned that the concepts behind Textual Communities should be promulgated. Firstly, we argue that scholars understand the reasoning behind the text/documents/entities division, with its insistence on the double aspect of the communicative act present in any textual document. Secondly, we argue that it should be possible for any textual scholar to make an edition, with a minimum of specialist computing and encoding knowledge and technical support. This requires that computing systems need to respond to the needs of scholars, rather than scholars restrict their editions to what computer systems can support. Thirdly, scholarly editing, perhaps more than any other area of the humanities, is uniquely positioned to profit from the social web. Scholars and readers may engage together in understanding documents, the texts they contain, and the complex histories of works which they compose. If we are able to move our discipline a small way in those directions, we will have done well. 44

Notes

[1] Although I am the sole author of this paper, at several points I speak of “we”. Many people have contributed to the thinking behind this paper, and to the Textual Communities system which seeks to implement that thinking. To name a few: the discussion of text, document and work is deeply indebted to many discussions over many years with Peter Shillingsburg, Paul Eggert, Hans Walter Gabler, David Parker and Barbara Bordalejo, among others (see, for example, [Shillingsburg 2007]; [Eggert 2009]; [Gabler 2007]; [Parker 1997]; and particularly the collection of essays in [Bordalejo 2013].This does not mean that any of these scholars agree with the definitions I offer: the best I can hope for, is that I have learned enough from them that they might disagree with me less vehemently now than they would have a few years ago. Here, “I” means “I”: those scholars are not responsible for my opinions. However, the creation of Textual Communities has been fully collaborative. The main implementation at Saskatchewan since 2010 has been the responsibility of Xiaohan Zhang, with some parts written by myself and Erin Szigaly. Throughout, we have consulted with Zeth Green (Birmingham) and Troy Griffitts (Münster), who have contributed key insights and questions.

The major projects using Textual Communities, named in footnote 17, have been an invaluable testing ground for Textual Communities. I am embarrassed for their long suffering over the last years, and grateful for their patience, support and encouragement. Finally, this article has benefitted significantly from the comments of many readers on a draft version posted on GoogleDocs in May 2015. I have indicated each place where I introduced a change suggested by one of these readers, and acknowledged the person. I also thank Torsten Schassan for instigating the discussion, and for his many corrections.

[2] So quoted by McGann #mcGann1983. For the source of the Morris citation, and the relationship of McGann's formulation to Morris's words, see [Noviskie 2013]. Noviskie traces the quotation to Sparling 1924.

[3] Compare the well-known FRBR (Functional Requirements for Bibliographic Records) Group 1 entities: "work, expression, manifestation, item" [IFLA 1998]; [Tillett 2004]. "Work" as I define it maps broadly to FRBR "work" (but might in different circumstances map to expression or manifestation); a document maps to an item. There is no equivalent in FRBR to "text" as I explain it. The system here presented extends, rather than replaces, FRBR. Particularly, this system enables works (and hence texts) to be seen as structured objects, readily susceptible to fine-grained manipulation, in ways that FRBR does not.

[4] This discussion of the stages of recognition of the text as a communicative act is indebted to the description by Barbara Bordalejo of the encoding system in the Prue Shaw edition of the Commedia, thus: "in this article, I use the phrase the 'text of the document' to refer to the sequence of marks present in the document, independently of whether these represent a complete, meaningful text. That is: the reader sees a sequence of letters, occurring in various places in relation to each other (perhaps between the lines or within the margins) and carrying various markings (perhaps underdottings or strikethroughs). These make up what I here refer to as the text of the document" [Bordalejo 2010].

[5] For example: an email by James Cummings to the TEI-L discussion on 5 February 2014 which speaks of encoding "both the interpretative `<text>` view and non-interpretative `<sourceDoc>` view" (http://permalink.gmane.org/gmane.text.tei.general/16892). Compare the prioritization of "document-based editing" over other kinds of editing argued by Hans Gabler [Gabler 2007], while noting that Gabler has always argued consistently that editors must also present the work (e.g., [Gabler 1984], [Gabler 1990].

[6] Among many contributions to the discussion of overlapping hierarchies in mark-up languages: see the original statement of "the OHCO thesis" in De Rose et al. [DeRose 1990] and its restatement and complication in Renear et al. [Renear 1993]. See too footnote 15.

[7] The author, while not formally a member of the workgroup ("TR9") on "Manuscripts and codicology" which was charged with drafting the chapter on representation of primary sources in the "P3" guidelines (first published in 1994), wrote most of the draft of what became Chapter 18 "Transcription of Primary Sources" in those guidelines (see http://www.tei-c.org/Vault/GL/P3/index.htm). This chapter persisted in revised form into the "P4" version, first published in 2002, before finally being replaced by Chapter 11 of the first "P5" version in 2007.

[8] To name a few: [O'Keefe 2006]; [Bornstein 2001]; and a whole European Society for Textual Scholarship conference on "Textual Scholarship and the Material Book: Comparative Approaches" in London in 2006. The emergence of the document as the locus of scholarly attention now has a name: "material philology". See too [Nichols 1990], and Matthew K. Driscoll "The Words on the Page", distilling talks given by him around 2005-2007 and available at http://www.driscoll.dk/docs/words.html.

[9] I know of only one substantial project that attempts this parallel encoding: the Goethe Faust project, [Brüning 2013].

[10] The Shelley-Godwin archive creators were fully aware of the arguments for encoding both "document" and "text", and canvas these in [Muñoz 2013], while confessing themselves unable to implement satisfactory encoding of both aspects.

[11] For the complexities of the *Visio Pauli* tradition see [Robinson 1996].

[12] I was aware, at an early stage, of the work of the Canonical Text Services (CTS) group, and studied their system closely while devising the system here devised [Blackwell 2014]. Briefly, this system is highly compatible with CTS, in that every CTS reference may also be expressed with no loss of information. However, the reverse is not true. This system includes completely hierarchical information for both document and communicative act, permitting full specification of both document space (a line within a page within a volume) and communicative act component (a word within a sentence within a paragraph within a chapter) to a degree that CTS does not enable. Further, CTS does not use the key/value pair architecture, nor does it specify the naming authority. It is in essence a labelling scheme, with some hierarchical elements, and relies on external index files to correlate (for example) text segments with the manuscripts in which they appear. For example: https://github.com/homermultitext/hmt-archive/blob/master/cite/collections/scholiainventory.csv includes the line "urn:cite:hmt:scholia.379","urn:cts:greekLit:tlg5026.msA.hmt:1.6","6th main scholia of Iliad 1","urn:cite:hmt:chsimg.VA012RN-0013@0.57311951,0.24827982,0.22451317,0.04644495","urn:cite:hmt:msA.12r". It appears that this links the "6th main scholia of Iliad 1" with the urn "urn:cts:greekLit:tlg5026.msA.hmt:1.6", and further with the manuscript urn:cite:hmt:msA.12r, presumably page 12r of "MsA". In contrast,

the system here described would yolk these statements together into a single URL, such as "…/document=MsA/folio=12r/entity=Scholia/Book=1/n=6", from which the full page and communicative act hierarchies could be deduced.

[13] The term "entity" is used here in preference to "work" for several reasons. Firstly, as the examples show, the term entity may be applied to a structured object of a communicative act at any level: a single line of *Hamlet* may be an entity; so too a single scene, an act, and the whole play itself are also entities. Secondly, the term "work", hotly contested in textual scholarship [Robinson 2013b], comes with many connotations which might not be helpful in understanding the system here proposed: "entity" has the advantage of neutrality. "Entity" is also familiar from FRBR, which is built on the categorization of relationships among "entities": distinct intellectual objects, analogous to the distinct components into which an act of communication may be structured.

[14] A description of the RDF ontology prepared in 2010, and an implementation of it, are available at http://www.textualcommunities.usask.ca/web/textual-community/wiki/-/wiki/Main/Historical+Documents

[15] As part of the move to MongoDB and JSON storage, Zeth Green, Xiaohan Zhang and myself reviewed how the document and entity hierarchies relate to the text formed from the collocation of the two hierarchies. We realized that it was possible, using a JSON-based architecture, to support not just two hierarchies for any text, but any number of hierarchies, thereby avoiding any difficulties with overlapping hierarchies. As of November 2016, we are still developing this new architecture. Recent articles by computer scientists demonstrate increasing awareness of the need to move beyond a "document paradigm", with its reliance on hierarchical content models, to systems which will natively support the multiply overlapping information schemes we find in actual texts and their physical instances: thus [Schmidt 2009] and, especially, [Schloen 2014]. At present and for the near future, however, XML in the TEI implementation remains crucial to our work, both because it keeps us close to a wide community of scholars working with digital editions and because of its sophisticated validation facilities.

[16] This paragraph describes the procedure used for designating document and entity parts in the first version of the Textual Community system. In the second version, we drastically simplified this: now, any TE/XML element with an "n" attribute becomes an element in either the document or entity hierarchy.

[17] The six projects are: at the University of Saskatchewan, the Canterbury Tales Project (with KUL, Belgium), led by myself and Barbara Bordalejo of KUL, 30,000 pages and 30 active transcribers; the John Donne Digital Prose Project, led by Brent Nelson, 1800 pages and 25 active transcribers; the Recipes Project, led by Lisa Smith, 1500 pages and 30 active transcribers; the Incantation Magic Project, led by Frank Klaassen, 400 pages and 10 active transcribers; at the University of Birmingham, UK, the Estoria de Espanna project, 3500 pages and 20 active transcribers; at City University New York, the Teseida Project, led by Bill Coleman and Edvige Agostinelli, 600 pages and four transcribers. Numerous other projects are also using the Textual Communities system, although it has not been publically launched.

[18] This use of `<pb/>`, `<cb/>` and `<lb/>` cannot cope with instances where the flow of lines on a page is disrupted by, for example, multi-line marginalia or annotations. We could use `<milestone/>` elements to mark out such instances.

[19] Brumfield does not mention two powerful systems for the creation of fully TEI-compliant XML documents in collaborative environments: TextGrid (www.textgrid.de/en/) and eLaborate (www.elaborate.huygens.knaw.nl/login). Both will support the making of the same complex TEI-XML documents as Textual Communities, but neither offers the same native support for both page-based and "text-based" transcription as does Textual Communities. Support for transcription by page, the best way to apportion transcription of full manuscripts among transcribers, is particularly crucial, and fully supported by Textual Communities.

[20] This sentence provoked a lively discussion in the Google Docs forum on the draft. Three commentators, Andrew Dunning, Hugh Cayless and Laurent Romary, questioned the need for the 'SA' condition. The nub of the problem is the expression of the 'SA' condition in Creative Commons and other 'copyleft' licenses, which insists that all further share-alike be under the same terms as the granting terms. This leads to a problem when a site wishes to mix together materials licencsed under different share-alike flavours: this cannot be done. Thus, although SA is conceived as a guarantee of continued open access, in practice it has become a very real restriction, inhibiting the free re-use we seek [Wiley 2007]. Accordingly, many recent open-access sites have dropped the SA requirement, and Textual Communities is likely to follow this lead.

[21] The second version of Textual Communities has abandoned the LifeRay environment here described: in practice, LifeRay has considerable difficulties, not least its vulnerability to "spam bots", which routinely dump spam within LifeRay documents.

Works Cited

Blackwell 2014 Blackwell, Christopher, Smith, Neel. "The Homer Multitext: Technically speaking", *The Homer Multitext*. http://homermultitext.blogspot.nl/2014/02/technically-speaking.html

Bordalejo 2010 Bordalejo, Barbara. “Appendix C: The Encoding System”, In Prue Shaw (ed.) *Dante Alighieri. Commedia. A Digital Edition*. Scholarly Digital Editions, Birmingham and Sismel, Florence (2010).

Bordalejo 2013 Bordalejo, Barbara (ed.) *Work, Text and Document in the Digital Age*, *Ecdotica*, 10 (2013).

Bornstein 2001 Bornstein, George. *Material Modernism: The Politics of the Page*. Cambridge University Press, Cambridge (2001).

Brüning 2013 Brüning, G, Henzel K., and Pravida, D. “Multiple Encoding in Genetic Editions: The Case of ‘Faust’”. *Journal of the Text Encoding Initiative* [Online]. URL: http://jtei.revues.org/697

Causer 2012 Causer, T., Tonra, J. and Wallace, V. “Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham”, *Literary and Linguistic Computing* 27, pp. 119-137 (2012).

DeRose 1990 De Rose, Steven, Durand, David, Mylonas, Elli, Renear, Allen. “What is Text, Really?”, *Journal of Computing in Higher Education* 1(2), pp. 3-26 (1990).

Eggert 2009 Eggert, Paul. *Securing the Past: Conservation in Art, Architecture and Literature*. Cambridge University Press, Cambridge (2009).

Eggert 2010 Eggert, Paul. “Text as Algorithm and as Process”. In W. McCarty (ed.) *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Open Book Publishers, Cambridge, pp. 183-202 (2010).

Gabler 1984 Gabler, Hans Walter. “The Synchrony and Diachrony of Texts: Practice and Theory of the Critical Edition of James Joyce’s Ulysses”. *Text*, 1, pp. 305–26 (1984).

Gabler 1990 Gabler, Hans Walter. “Textual Studies and Criticism”, *The Library Chronicle*, The University of Texas at Austin, pp. 151–65 (1990).

Gabler 2007 Gabler, Hans Walter. “The Primacy of the Document in Editing”, *Ecdotica*, 4, pp. 197–207 (2007).

IFLA 1998 *Functional Requirements for Bibliographic Records*. IFLA Series on Bibliographic Control. Munich, K. G. Saur (1998).

Kahn 2006 Kahn, Robert and Wilensky, Robert. “A Framework for Distributed Digital Object Services”, *International Journal on Digital Libraries* 6(2), pp. 115–123 (2006).

Maryland 2014 Maryland Institute for Technology in the Humanities, et al. *Shelley-Godwin archive*. http://shelleygodwinarchive.org/

McGann 1983 McGann, Jerome J. *A Critique of Modern Textual Criticism*. Chicago, Chicago University Press (1983).

McKenzie 1999 McKenzie, Donald F. *Bibliography and the Sociology of Texts (The Panizzi Lectures, 1985)*. Cambridge University Press, Cambridge (1999).

Muñoz 2013 Muñoz, Trevor, Viglianti, Raffaele, Fraistat, Neil. “Texts and Documents: new challenges for TEI interchange and the possibilities for participatory archives”, *Abstracts for TEI Members conference* (2013). http://www.tei-c.org/Vault/MembersMeetings/2013/wp-content/uploads/2013/09/book-abstracts.pdf#page=99

Möller 2005 Möller, Erik. “Creative Commons -NC Licenses Considered Harmful”. http://www.kuro5hin.org/story/2005/9/11/16331/0655

Nichols 1990 Nichols, Stephen (ed.) *“The new philology”, Special issue of Speculum: A Journal of Medieval Studies*, LXV (1990).

Noviskie 2013 Noviskie, Bethany. *Resistance in the materials*. http://nowviskie.org/2013/resistance-in-the-materials/

O’Keefe 2006 O’Keefe, Katherine O’Brien. *Visible Song: Transitional Literacy in Old English Verse*. Cambridge University Press, Cambridge (2006).

Parker 1997 Parker, David C. *The Living Text of the Gospels*. Cambridge University Press, Cambridge (1997).

Pierazzo 2011 Pierazzo, Elena. “A Rationale of Digital Documentary editions”, *Literary and Linguistic Computing*, 26 pp. 463-477 (2011).

Renear 1993 Renear, Allen, Mylonas, Elli, Durand, David. *Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies*, Brown University. http://www.stg.brown.edu/resources/stg/monographs/ohco.html

Robinson 1996 Robinson, Peter M. W. “Is there a text in these variants?”, In R. Finneran (ed.) *The Literary Text in the Digital Age*. University of Michigan Press, Ann Arbor, pp. 99-115 (1996).

Robinson 2004 Robinson, Peter M. W. (ed.) *Geoffrey Chaucer. The Miller’s Tale on CD-ROM*. Scholarly Digital Editions, Leicester (2004).

Robinson 2013a Robinson, Peter M. W. “Towards A Theory of Digital Editions”, *Variants*, 10, pp. 105-132 (2013).

Robinson 2013b Robinson, Peter M. W. “The Concept of the Work in the Digital Age”, In Barbara, Bordalejo (ed.) *Work, Text and Document in the Digital Age*, *Ecdotica*, 10, pp. 13-41 (2013).

Robinson forthcoming Robinson, Peter M. W. “The Digital Revolution in Scholarly Editing”, *Ars Edendi Lecture Series*, IV, Stockholm University, Stockholm.

Schloen 2014 Schloen, David, Schloen, Sandra. “Beyond Gutenberg: Transcending the Document Paradigm in Digital Humanities”, *Digital Humanities Quarterly*, 8(4) (2014).

Schmidt 2009 Schmidt, D. and Colomb, R. “A Data Structure for Representing Multi-version Texts Online”, *International Journal of Human-Computer Studies*, 67, pp. 497-514 (2009).

Shaw 2010 *Shaw, Prue. Dante Alighieri. Commedia. A Digital Edition*. Scholarly Digital Editions, Birmingham and Sismel, Florence (2010).

Shillingsburg 2007 Shillingsburg, Peter. *From Gutenberg to Google*. Cambridge University Press, Cambridge (2007).

Sparling 1924 Sparling, Henry H. *The Kelmscott Press and William Morris, Master Craftsman*. London (1924).

Sutherland 2010 Sutherland, Kathryn (ed.) *Jane Austen’s Fiction Manuscripts: A Digital Edition*. http://www.janeausten.ac.uk.

Tillett 2004 Tillett, Barbara. *FRBR: A Conceptual Model for the Bibliographic Universe*. Library of Congress Cataloging Distribution Service (2004).

Trinity College Dublin 2014 *Letters of 1916: Creating History*. http://dh.tcd.ie/letters1916/

Wiley 2007 Wiley, David. “Noncommercial Isn’t the Problem, ShareAlike Is”, *Open Content*, 17 July 2007. http://opencontent.org/blog/archives/347 (Accessed 25 June 2015).