Encoding standards for the electronic edition

Lou Burnard, Oxford

Prispevek predstavi temeljne pojme in predstave, na katerih slonijo *Smernice za elektronsko kodiranje in izmenjavo besedil* konzorcija Text Encoding Initiative (TEI) ter še zlasti pristop, s katerim te *Smernice* rešujejo tekstnokritična vprašanja. Razprava pretresa možne prednosti digitalno kodiranih besedil, zlasti na področjih, kot sta korpusno jezikoslovje in tradicionalna filologija, in opiše čedalje pomembnejšo vlogo, ki jo na teh področjih zavzemajo digitalni viri in orodja. Sklepni del zagovarja stališče, da imata kodiranje besedil in digitalizacija širše metodološke implikacije in porajata teme, ki so bistvenega pomena za humanistične cilje in prizadevanja.

1. Abstract

As the dust settles on the end-of-century debate about whether or not the e-book will ever replace the codex (or c-book), a few simple and obvious truths remain. Firstly, barring some kind of cataclysm, to return to a world in which digital technologies are at the margins of scholarship is almost inconceivable. Whatever else it may become, the e-book, and its associated support system, the internet, now constitute the main communications channel for the written word (and much else besides) in the academic world. Secondly, there are some specific aspects of digital techniques which radically change the cost/benefit balance in making those communications, even without considering the potential for change in the social, industrial, and political systems underpinning our present-day notions of what constitutes "publication". And thirdly, for many practitioners, it is this (possibly subversive) change in the definition of what 'editing' and an 'edition' are which constitutes the most exciting opportunity offered by a world of digitized cultural resources. In this paper, first presented at the conference on *Scholarly editions in the electronic medium* held in Ljubljana, June 2004, I focus on the way in which digital editions actually instantiate, and thus problematize, very traditional philological notions.

As a vehicle for presenting this idea, I summarize some key aspects of the model of texts and textual endeavours which underly the Recommendations of the Text Encoding Initiative, in particular the primacy they attach to a view of texts as meaningful constructs which have a place in time and space, rather than as visual artefacts whose meanings are purely contingent. I emphasize the extraordinary scope of the *TEI Guidelines*, which were produced by and for experts drawn from several distinct academic communities (librarians and computer scientists, philologists, historians, linguists...), suggesting both that this renaissance spirit anticipates the joyful eclecticism of today's digital media, and also that those new technologies which emphasize the fragmentation and aggregation of digital resources (collaborative research, e-research, grid technologies, web services...) are really not so far removed from it.

Finally I discuss a specific project which applies the *TEI Guidelines* to a traditional editorial ambition: the production of a new online edition of an existing and much-edited medieval text, the *Ancrene Wisse* (see http://www.tei-c.org.uk/Projects/EETS/). This case study demonstrates how the TEI scheme permits us to capture many forms of the text, as page images, as diplomatically transcribed pages, as a 'synthetic' or edited text, and as a modern translation. These forms are closely integrated in a web site presentation, using open standards such as XSLT and XML. Substantial numbers of other such texts could also be integrated, together with detailed metadata descriptions, to provide an online corpus of material, for use perhaps in linguistic applications far removed from the original goal of the philologist concerned with the specific edition. Yet the work of that philologist, drawing as it does on an awareness of context and background, would also be enhanced by the availability of these resources.

2. New wine or new bottles?

> And almost thence my nature is subdued
> To what it works in, like the dyer's hand:
> Shak. Son. 111

Technology is not neutral. As Shakespeare reminds us, there is a dyer's hand effect. Our expectations of what is achievable in any domain, and thus the goals we set ourselves, are inevitably affected by our knowledge of what is technically feasible. It is in this way that the myriad small quantitative changes (in operational efficiency, in ease of use, in accessibility etc.) following our decision to 'go digital' when editing a text gradually approximate to a qualitative change. Our enterprise is subtly transformed, not as a consequence of any sinister dehumanizing consequences of the technology, but simply because it enables us to achieve something recognisably the same as hitherto, but to do so more efficiently, and more effectively.

I would like to suggest two specific examples of these rather general principles: the first focuses on the extent to which the art or science of corpus linguistics approximates to providing us with a new view of language; the second on the extent to which the craft or science of electronic editing approximates to a new view of texts and textuality.

2.1. Corpus linguistics

Corpus-based linguistics is, in Hoey's memorable phrase, 'not a branch of linguistics: but a route into linguistics'. This is not the place to rehearse the history and principles by which the formal study of language-in-use, of patterns of lexis, and other linguistic phenomena, has become a major component of all aspects of the field of linguistics, nor to revisit old controversies about the relative merits of evidence-based and of theory-based approaches to the study of language. The interested reader may find this in any of the several excellent introductions to this no-longer emergent discipline now available.[1] Twenty years ago, this was an esoteric research area; now (in the UK at least) it is a component of the National Curriculum.

Its relevance to our present discussion is the simple observation that it represents an area of study which would be simply impossible without the assistance of computers. We are unsurprised to hear of major scientific areas that have been irreversibly changed by the advent of cheap computing power; the success of corpus linguistics reminds us that this technology has also revolutionized the practice of lexicography, language pedagogy, and even the basics of linguistic theory, and theories of cognition.[2]

[1] McEnery and Wilson, Introduction to Corpus Linguistics (Edinburgh University Press, 1996) is a good example, also supported by a useful web site http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm.

[2] For a recent, and impressive, example of the contributions made by the corpus approach in linguistic theory see Hoey, Lexical Priming, (Routledge, 2005).

2.2. Electronic editing

If this is true of linguistics, it is also true of the production of books in general, and of scholarly books in particular. Is there, in fact, anything new about the 'e-book'? It is certainly loudly claimed that the proliferation of digitized resources offers new kinds of evidence. Such resources are available to more, and more varied, readers; they also are in themselves more, and more varied. This expansion of access and of materials has some evident effects (or are they causes?) in the twin pressures of greater sensitivity to multiple cultures, and of the decanonization of specific cultural norms.

There is another effect following on from what I have elsewhere characterized as the rise of a digital demotic: as media converge on a single digital format, there is a risk that in the erosion of difference, critical sensibilities and cultural awareness also disappear. As well-informed academics, we may wring our hands and bemoan the fact that in the age of the web, never before have so many known so little about so much. But it is our responsibility to counter this trend, by taking advantage of the opportunities afforded by new media and new technologies, rather than railing against the disneyfication of cultural values we may perceive around us. As with the technology itself, this is another case where many small quantitative changes approximate a qualitative one.

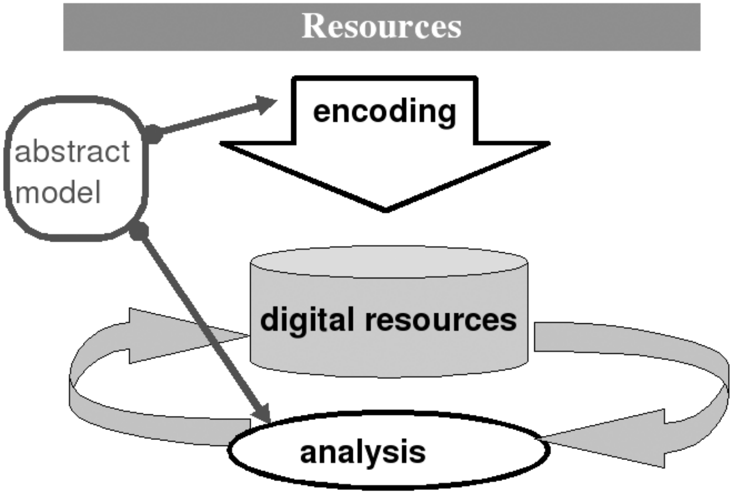
What are the major components of a traditional humanities masters degree? In most European universities, I believe, they focus above all on methodological and hermeneutic issues: questions of interpretation and the management of accumulated interpretations predominate. The traditional demonstration that one has achieved mastery of these skills is twofold: the production of a new critical edition, and its successful defence against interrogation by ones peers. While not wishing to subvert the importance of these skills, I would like to argue that they need to be reapprised and perhaps rediscovered in the new digital context. In place of the critical edition, perhaps we need to teach more about the production of the *uncritical* edition: the diplomatic transcription of a complex textual source. The tools now at our disposal allow us to offer at least the appearance of unmediated access to the complexities of a textual tradition hitherto shaped by the decisions of the traditional editor. Although each reading in such a tradition is, of necessity, culturally and chronologically determined, the uncritical edition allows us to compare and contrast their effects, to amass the evidence and its selective re-presentations, to allow each such reading its own space, within its own context.

Moreover, the insights of critical editing and traditional philology are sorely needed on the internet, at least as much as they are needed in the library. An awareness of the fluidity of a text and of the unreliability of the witnesses to it is as essential a skill for today's web surfer as is an awareness of the need to seek independent corroboration for all assertions, and a readiness to investigate the motivations and context of those making them. Such awareness, such readiness, is what characterizes at its best the humanistic scholarly tradition. This potent and fruitful synergy of semiotics, textuality, and hermeneutics is what we have to offer: we should not undervalue it.

Nevertheless, there are qualitative differences in the way we interact with digital and with non-digital resources. The fugitive physicality of the digital text favours a decentred, non-linear, fragmented, and associative mode of cognition, which (however hard we may assert the contrary) is hard to bring to bear when dealing with traditional physical books. The immense differences of scale between the texts which can be held on one's bookshelf and on one's hard disk complicate our sense of what is an appropriate contextual norm for any textual assertion. How puny one's awareness of the language of a single author's complete works seems when compared to the textual universe available in digital form! Yet, given an elementary grasp of statistics, how enriching is a perception of the relation between decontextualized instances of language use and the norms of the context from which they come. The notions of greatness and moral value associated with great art can perhaps be enriched by a sense of their statistical deviance, of the way they play with and problematize cultural norms and expectations, for example.

The plasticity of format and presentation associated with digital resources, as I suggested above, seems to approximate a qualitative difference. The ability to merge media of different types (sound, image, transcription, metadata…) into a single resource encourages us to think more about what it is that such resources have in common and less about what contingently distinguishes them. When everything can be reduced to a bit-stream, what causes us to attach more importance to some bit-streams than to others? The traditional answer is to point to the fact that some bitstreams are not transparently significant but require an explication, indeed that it is only in that explication that significance emerges from them. Explication or interpretation is thus what confers value, and, as Derrida, citing Montaigne, remarks 'We need to interpret interpretations more than to interpret things'.

The relevance of this to our present discussion is that the process of digitization is precisely a process of capturing a set of interpretations. Digitization reifies a particular reading or set of readings for a text to the exclusion of others. Resources must be decoded (interpreted) before they can be encoded. Both the decoding of a resource (carried out in order to represent it) and the subsequent encoding of that reading of the resource necessarily imply selection of features according to some particular agenda or formal model. However, once in digital form, resources can be enriched by comparison with others, by repeated analyses or interpretations in a virtuous hermeneutic circle. The following figure attempts to summarize this process:



The scientific paradigm tries to free itself from the complications inherent in the thought that all observations (and hence explanations) are conditioned to greater or lesser extent by the person observing. But the paradox of Schrödinger's cat is one with which the humanities have long been familiar: the observer effect, however novel or deprecated in the sciences, is one which the humanities have long contended with. In the same way, computing systems are typically used in the humanities for the manipulation, storage, and preservation of symbols in abstract semiotic systems, rather than as machines for mathemetical calculation.

The digital edition should thus be seen as a repository within which encoded messages about and representations of a text or other resource can productively coexist without loss of integrity. Scholarship itself depends on a continuity — it is not enough simply to preserve the encoded text; there must also be a continuity of comprehension. The entry ticket for the permanence afforded by indefinite migration from one storage medium to another is the recognition that it is only media-independent features of the encoding which can be so preserved.

In these respects, I suggest that digital editions actually instantiate, and thus problematize, precisely those traditional philological notions of accuracy with respect to a source and understanding of the cultural context from which that source derives. In other words, keeping faith with the hermeneutic systems associated with a source is an inextricable part of its retransmission, irrespective of the media of transmission.[3]

3. The tools of the trade

Why should the philologist bother to learn about the technologies underlying the digital edition, rather than simply making use of them? One does not need to know how typewriters are constructed in order to write a bestseller. But technologies like XML and the TEI are not passive instruments to be used off the shelf in only one pre-packaged way. They are enabling, even empowering, technologies, which return control of resources to their owners, instead of locking them away into blackbox proprietary systems.

An understanding of XML therefore should matter greatly to everyone who has textual data to process, particularly if they want to share that data with others now, or in the future, or if they want to take advantage of the immense quantity of other people's software being produced. XML allows us (almost for the first time) to represent textual structures, metadata, and multiple analyses within a single encoding framework. This insight is already transforming the way business information systems are designed and implemented, as a glance at the trade press will confirm. However, XML in its TEI application also provides us with a flexible system appropriate to the handling of complex interpretive data from a humanistic and scholarly perspective.

Adopting XML as a standard facilitates the integration and sharing of resources which characterize the humanities tradition at its best. Learning how it works

[3] For a range of essays on the nature and effect of 'Electronic Textual Editing', see the volume of that title, edited by Unsworth, O'Keefe, and Burnard, currently in press with the Modern Language Association, and available in preprint form from http://www.tei-c.org/Activities/ETE/Preview/.

should be regarded therefore not as a peripheral activity like learning to type, but rather as an essential part of the discipline, like learning to write.

3.1. What is XML?

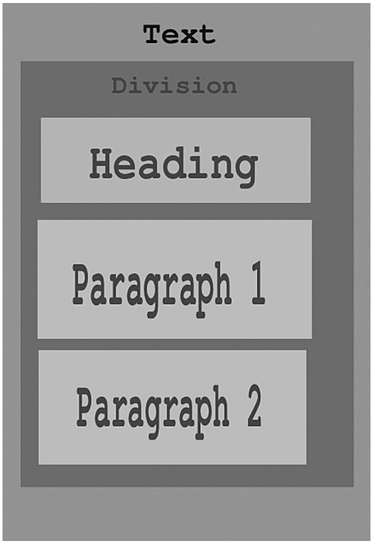
An article of this length is not the place from which to learn the details of XML and its TEI implementation.[4] However, it may be helpful for the complete novice to mention some of the salient features of this way of thinking about what a document is.

In an XML document, information is represented by means of identifiable objects, called elements. The document itself is such an object, and it contains within it nested occurrences of other objects of the same or other types. Objects of any type may also bear descriptive named attributes. A very simple document, like this paper, might consist of a `<text>` element, containing a number of `<division>` elements. Each `<division>` element might begin with an optional `<heading>`, and then contain a number of `<paragraph>` elements. Each `<paragraph>` might contain just plain text, and might have an attribute **lastChange** specifying the time and date it was last changed.

An important characteristic of an XML document is that its structure can be described by means of a textual grammar (technically known as a schema). A schema specifies both what elements exist, and how they can meaningfully be combined (for example, that a `<division>` element may contain a `<heading>` only if it precedes a `<paragraph>`). Depending on the specific schema language employed, rules may also be specified which constrain the legal values for attributes, for example to specify exactly the format for representing dates and times.

Every XML document is modelled as a single hierarchy of elements, which means that it can easily be represented as a linearised sequence with labelled brackets. So, for example, by marking the start and end of texts, divisions, headings, and paragraphs, the nested structure in this figure:

[4] Fortunately there is no shortage of introductory material on such topics on the Web: the TEI's own 'Gentle Guide' is at http://www.tei-c.org/P4X/SG.html.



can be represented without loss of information by the following linear sequence:

```
<text><division>
<heading>Heading</heading>
<para>Para 1</para>
<para>Para 2</para>
</division></text>
```

This tree-based model of textual structure has many advantages from both conceptual and processing points of view. However, it is important to note that one can often define more than one hierarchic structure within a text and moreover that the components of such structures rarely map tidily onto one another. A codex may be seen as a hierarchic structure of gatherings containing leaves, each comprising a recto page and a verso page. A page may contain many paragraphs, but a single paragraph may start on one page and finish on the next. We cannot therefore easily nest a paragraph hierarchy within a page hierarchy. Similarly, in a verse drama, because we can expect to find verse lines which are divided between speakers, we cannot easily combine speaker and metrical structures. A number of techniques have been proposed for the representation of such structures in XML, of which the majority depend on the language's ability to represent links between one point of a structure and another, thus making it possible to define one or more independent hierarchies while at the same time representing their points of alignment, both with each other and with a single segmentation of the text being annotated.

XML markup thus provides a unified way of representing individual tokens in a text, the structural units in which they occur, arbitrary regions of text, and links, correspondences, or alignment amongst these elements. These features of a text can also be formally verified against a textual grammar. XML markup can be verified in such a way as to ensure that it reflects the meaning of your data, not its appearance.

3. 2. What is the TEI?

In November 1987, representatives from about forty academic institutions and projects converged on Vassar College, Poughkeepsie in upstate New York. They shared a common vision: the transformation of substantial amounts of the world's literature into computer-readable form; but what had brought them together was the lack of agreed international standards in the production of what was at that time called 'machine-readable text'. Over two wintry days, they considered the cacophony of different practices emerging in the service of that common vision at institutions already existing across the US, across Europe, and in Japan, but scattered and fragmented in those pre-World Wide Web days. The Internet as a social phenomenon was approaching adolescence: it had not yet established itself outside the world of academic research, and the technical standards needed to move it forward were still only dimly perceived.

Following that conference,[5] the Text Encoding Initiative (TEI) was formed as an international research project, with funding from the US National Endowment for the Humanities, the European Union, the Canadian Social Science Research Council, and the Andrew W. Mellon Foundation. The TEI was jointly sponsored by three established international professional associations, which established a small management committee, and appointed two 'editors' to co-ordinate the enthusiastic participation of more than a hundred scholars worldwide. Its remit was to attempt a complete definition of current practice and to produce recommendations or Guidelines for the creation and usage of electronic texts in key linguistic and literary disciplines. The first research phases of the TEI came to an end in 1994 with the publication of TEI P3, the 'big green books' which over the next few years were to become the reference standard for the building of the digital library.[6]

[5] For a report on the outcomes of the Poughkeepsie Conference, see 'Report of Workshop on Text Encoding Guidelines' in Literary & Linguistic Computing, 3 (1988).

[6] For history and background on the TEI, see the website at http://www.tei-c.org.

At the start of the current century, the TEI re-established itself as a membership consortium, jointly hosted by two US and two European Universities, and sponsored a first major revision of the TEI Guidelines. This edition, known as TEI P4, was a 'maintenance release', bringing the Guidelines up to date with changes in the technical infrastructure — most notably in the use of the W3C's Extensible Markup Language (XML) as its means of expression[7] rather than the ISO standard SGML. TEI P4 was published in 2002, under the imprint of the University of Virginia Press, and forms the current reference standard.

However, nothing written in digital form is ever really finished. Since 2002, work has been proceeding on the next major revision of the TEI Guidelines, to be known as TEI P5, which will include far more substantive changes than were needed for P4.[8]

The goals of the TEI were to facilitate better interchange and integration of scholarly textual data, in all languages, and from all periods. As such the TEI had two inherently contradictory objectives: on the one hand, to deliver 'guidance for the perplexed' by suggesting *what* textual features should be encoded in a document; on the other, to provide assistance for the specialist by suggesting *how* any particular set of textual phenomena might be encoded. Its aim was thus to be both a user-driven codification of existing best practice, and also a loose framework into which unpredictable extensions might be fitted. Its recommendations thus cover both generic text structures and some highly specific areas based on (but not limited by) existing practice.

The original scope of the TEI was encyclopaedic, embracing not just the basic structural and functional components of running text but also diplomatic transcription of non-digital textual (and aural) sources, images, and annotation thereof; links, correspondence, and alignment of encoded objects; data-like objects such as dates, times, places, persons, events (what is now termed 'named entity recognition'); meta-textual annotations (correction, deletion, etc); linguistic analysis of all kinds and at all levels; contextual metadata of all kind. No-one requires all of this, yet all of it is required by someone. Considerable effort and ingenuity was therefore invested in ways of making it possible to derive multiple views (DTDs or

[7] This was particularly appropriate in that one of the editors of the XML standard, Michael Sperberg-McQueen, had also served as editor of the original TEI Guidelines.

[8] A preliminary release of TEI P5 appeared in January 2005: see http://www.tei-c.org/P5 for its current status.

schemas) from the enormous set of textual categories or distinctions identified by the Guidelines, which (as of TEI P4) documented 362 distinct XML elements, 95 attributes, and 88 classes, grouped into 24 distinct modules of various kinds. To demonstrate that this polymorphic architecture actually worked, the TEI editors used it to produce TEI Lite, one specific customization which turned out to have a life of its own.

TEI Lite was intended to meet three goals. The first was to provide a subset of the TEI Recommendations which would be adequate to the needs of most likely users of the whole of the TEI most of the time. The second goal was to provide a subset of elements rich enough to support an authoring environment for the production of online documentation such as the Guidelines themselves. And, as already mentioned, TEI Lite was conceived of as a practical demonstration of the customization facilities of the TEI scheme — and therefore had to provide some features which were *not* otherwise included in it.

Despite the protestations of the TEI editors and others, however, TEI Lite was occasionally perceived as being *the* recommended introduction to the TEI scheme for all purposes. Widely translated, its manual came to form the basis of many introductory tutorials and workshops,[9] as well as to define the practice of many major encoding projects, particularly in the digital library community. This was a mixed blessing: TEI Lite is (inevitably) lacking in elements some projects will consider essential, and (even more inevitably) over-supplied with elements other projects will never wish to use. Because the customization methods by which TEI Lite was constructed remain somewhat arcane to the non-SGML specialist, there was a tendency for projects to edit TEI Lite itself in nonstandard and unpredictable ways to produce 'TEI Lite like' schemas, thus seriously compromising the interchangeability of their resources. More worryingly, it came to be perceived in some quarters as the TEI's final word on what 'text' *really* is: though any careful reading of the TEI Guidelines proper would show the extent to which the TEI tries to problematize this notion, by emphasizing the relativity of the shared assumptions and priorities about the digital agenda which underly its suggestions. Those priorities, specifically the focus on content and function (rather than presentation), and the identification of generic solutions (rather than application-specific ones) have stood the test of time. But they are not necessarily universal.

[9] See http://www.tei-c.org/Lite/.

All texts are alike, in some sense; yet every text is different. Because it needed to be able to cope with the full variety of texts and textual readings, and the whole range of scholarly endeavours, the TEI system was designed in a modular way. Rather than define a single monolithic schema, the TEI defines a number of schema modules which can be combined in a controlled manner. Each module contains definitions for a set of elements and attributes, each of which is also a member of one or more classes. Elements may refer to each other indirectly by means of their class membership. This allows the system designer to modify particular components of a module, or to add new components to it, without disturbing the rest of the structure. The technical details of the mechanisms used to implement this architecture are described elsewhere;[10] the key point is that the system can be used to develop a schema which contains all and only the elements and attributes needed to support the specific needs of a given text-creation project without overly compromising the interchangeability of the project's deliverables.

4. A case study

In conclusion, I discuss a small scale TEI implementation developed for the Early English Text Society (EETS).[11] This is one of the oldest-established British learned societies in the field of medieval studies, which has been publishing well-respected primary editions of medieval materials since its founding in the mid-19th century. Consequently, the EETS has accumulated a long list of print editions of scarce resources, which it would be of considerable interest retrospectively to digitize.

In 2003, Dr Bella Millett was awarded a small AHRB research grant to investigate the feasibility of transforming a typical print EETS edition into a digital resource. The text chosen, the *Ancrene Wisse,* is a well known Middle English 'rule' or 'guide' for female recluses, composed in the West Midlands in the early thirteenth century. Her forthcoming edition of this text for EETS (based on the un-

[10] For TEI P4, see for example 'Rolling Your Own with the TEI' in Information Services and Use vol 13 no 2 (Amsterdam, IOS Press, 1993); for P5, see Sebastian Rahtz,'Converting to Schema: the TEI and RelaxNG'. Available from http://www.idealliance.org/papers/dx_xmle02/papers/03-03-08/03-03-08.html. Paper presented at XML Europe 2002, Barcelona, May 2002; see also Burnard and Rahtz, 'RelaxNG with Son of ODD', Available from http://www.mulberrytech.com/Extreme/Proceedings/html/2004/Burnard01/EML2004Burnard01-toc.html. Paper presented at Extreme Markup Languages, Montréal, August 2004.

[11] See http://www.tei-c.org/Projects/EETS/.

completed edition by Eric Dobson, with a glossary by Richard Dance) was adopted as a starting point for the project. This edition contained introductory notes on the manuscripts, and a traditional apparatus criticus derived from a full collation, but no images or transcription of the manuscripts. The *Ancrene Wisse* is a substantial prose text, surviving in seventeen different manuscripts; for this pilot, we worked only with the preface, for which digitized page images of four early manuscripts were obtained and transcribed. The material in the original edition was revised and further augmented by these images, and also by a diplomatic transcription of the four manuscript versions, as well as a translation of the edited copy text.

All of this material, it was agreed, should be integrated for delivery over the web, thus providing much more than the typical scholarly edition in both quantitative and qualitative terms. The material would all be stored as TEI XML, but converted to HTML for the convenience of current generation web browsers.

The components of the edition are organized and encoded in such a way as to facilitate the automatic definition and processing of links between the various components. More specifically, it is possible to present on the screen corresponding parts of a text in different manuscript versions, in both page images and diplomatic transcript, and in the edited text. Similarly, it is necessary to align text and translation, to associate entries in the bibliography or notes with discussion elsewhere in the text. All this linking was represented using the ID/IDREF mechanism of XML, which meant that we had to define XML elements for every text component that might potentially be aligned in this way.

The structure proposed for this EETS pilot e-edition may be summarised as follows. The entire edition is tagged as a single `<TEI.2>` element, containing (as usual) a `<teiHeader>` followed by a `<text>` element. The `<teiHeader>` contains descriptive metadata for the entire edition (including details of any codes used in more than one of its components such as manuscript hand identifiers). The `<text>` element groups together, at the top level, the following:

- front matter for the whole edition, tagged `<front>`, containing all the introductory material;
- a `<group>` element, containing a number of `<text>` elements, each containing a distinct version of the text in question, possibly with its own front and back matter;
- back matter for the whole edition, tagged `<back>`, containing sections of bibliography, analytic notes, glossary, and index.

By 'text' above, we meant a number of distinct things. Specifically, we wished to distinguish the following types:

edited
An edited text
translated
A translated text
mss
A group of transcriptions
text-trans
An aligned virtual text

By 'virtual' text, we mean here a text that is to be generated automatically from others. For example, we wished to display text and translation as a single text, aligned at paragraph breaks; since the whole document was generated as a series of static web pages, we needed to include some element to indicate where this aligned version was to be generated.

Within each text component, the `<div>` element was used to mark further subdivisions. The TEI `<divGen>` element was used to mark where 'virtual' subdivisons are to be generated, for example for divisions comprising links to page images, or to manuscript descriptions.

To simplify the encoding of the diplomatic transcriptions, we made a number of small modifications. We removed a large number of elements which we had no need of. We introduced new elements, chiefly convenience shortcuts such as `<lat>` (for `<foreign lang="lat">`) or `<fr>` (for `<foreign lang="fr">`) etc. We also introduced specialised pointing elements such as `<pp>`, used in the same way as the standard `<ptr>` element, but with the added constraint that its target should be a <p> element. And we introduced an element `<page>`, modelled on the TEI `<ab>` element, to delimit all the text on a given page.

From these modifications, we generated a project-specific DTD using the TEI Pizza chef tool[12] and used it to validate the XML components of the new edition. Note that, with these additional elements, we were able to represent both conflicting hierarchic views of the text: in the transcript texts, each page is a block

[12] http://www.tei-c.org/pizza.html. At TEI P5, a much improved TEI schema builder called Roma is available, which amongst other enhancements facilitates the generation of project-specific documentation.

(<page>), with the boundaries of paragraphs represented by empty pointing elements (<pp>); in the edited texts, each paragraph is a single block (<p>), with the page boundaries represented by empty pointing elements (<pb>). This symmetry of treatment turned out to simplify considerably the creation of XSLT stylesheets for the task of rendering combined views of the whole set of texts as static HTML pages, which was our remaining task.

Constraints of time and funding did not enable us to treat all the components of this edition as fully as might have been desired: for example, we were not able to include full TEI-conformant descriptions of the manuscripts themselves, nor were we able to transform the glossary provided into an integrated lexicon, or to provide more than a rudimentary text searching component. Nevertheless, the exercise demonstrated the overall viability of the TEI approach as a means of producing a sophisticated electronic product, capable of supporting the demands of traditional philology.

5. Conclusions

Far from being peripheral or in opposition to the humanistic endeavour, text encoding and digitization are central to it. Text encoding is the best technique at our disposal for ensuring that the hermeneutic circle continues to turn, that our cultural traditions endure. The TEI scheme provides a solid basis for applying such techniques in the service of traditional philological insights, as well as enabling newer more exploratory applications for them.

Standardi kodiranja za elektronske izdaje

Lou Burnard, Oxford

Potem ko se je polegel prah po polemiki ob koncu stoletja, ali bo e-knjiga kdaj nadomestila kodeks (ali k-knjigo), je ostalo nekaj preprostih in očitnih resnic. Prvič: vrnitev v svet, v katerem so bile digitalne tehnologije na robu znanosti, je skoraj nepredstavljiva, razen če ne bo prišlo do kakšne katastrofe. Naj se zgodi kar koli drugega – e-knjiga in z njo povezan podporni sistem, medmrežje, sestavljata glavni komunikacijski kanal za pisano besedo (in poleg tega še za veliko drugega) v akademskem svetu. Drugič: obstajajo posebni vidiki digitalnih tehnik, ki pri ustvarjanju teh komunikacij radikalno spreminjajo ravnotežje med stroškom in koristjo, in to celo brez upoštevanja možnosti za spremembo v družbenih, industrijskih in političnih sistemih, ki podpirajo naše dandanašnje predstave o tem, kaj predstavlja »publikacijo«. In tretjič: za veliko ljudi, ki se s tem ukvarjajo, je prav ta (mogoče subverzivna) sprememba definicije, kaj sta »izdajanje« in »izdaja«, tisto, kar predstavlja najbolj vznemirljivo priložnost, ki jo ponuja svet digitaliziranih kulturnih virov. V tem predavanju, prvič predstavljenem na konferenci o *Znanstvenih izdajah v elektronskem mediju* junija 2004 v Ljubljani, sem pozornost osredinil na to, kako digitalne izdaje pravzaprav udejanjajo in hkrati problematizirajo nekatere zelo tradicionalne filološke pojme.

Kot vodilo pri predstavitvi te ideje sem poskusil povzeti nekaj ključnih vidikov besedilnega modela in tekstoloških prizadevanj, na katerih slonijo *Smernice* konzorcija Text Encoding Initiative, še posebej pomembnosti, ki jo pripisujejo pogledu na besedila kot pomenljive konstrukte, ki imajo svoje mesto v času in prostoru, ne pa pogledu na besedila kot vizualne artefakte, katerih pomeni bi bili zgolj naključni. Prav tako sem nakazal izjemen obseg delovanja *Smernic TEI,* ki so jih za strokovnjake izdelali strokovnjaki. Ti so bili pritegnjeni iz več specifičnih akademskih skupnosti (knjižničarji in računalniški znanstveniki, filologi, zgodovinarji, lingvisti …), kar napeljuje k temu, da ta renesančni duh anticipira veseli eklekticizem današnjih digitalnih medijev, pa tudi k temu, da nove tehnologije, ki poudarjajo fragmentacijo in agregacijo digitalnih virov (raziskovanje v sodelovanju, e-raziskovanje, medmrežne tehnologije, spletne storitve …), pravzaprav niso daleč od njega.

Končno sem obravnaval poseben projekt, ki aplicira *Smernice TEI* na tradicionalno izdajateljsko ambicijo: pripraviti novo medmrežno izdajo obstoječega in večkrat izdanega srednjeveškega besedila *Ancrene Wisse* (glej http://www.tei-c.org.uk/Projects/EETS/). Obravnava tega primera pokaže, kako nam shema TEI omogoča, da zajamemo več oblik besedila, kot so slike strani, diplomatični prepisi strani, »sintetična« ali urejena besedila in moderni prevodi. Te oblike so tesno povezane v spletno prezentacijo, ki sloni na uporabi odprtih standardov, kakor sta XSLT in XML. Na enak način bi lahko povezali v elektronsko prezentacijo veliko število podobnih besedil, vključno z njihovimi podrobnimi metapodatkovnimi opisi, da bi tako ustvarili medmrežni korpus gradiva, uporabnega morda za jezikoslovne aplikacije, že močno oddaljene od prvotnega namena filologov, ki pripravljajo specifično kritično izdajo. In vendarle bi bilo delo takega filologa, ki je zadolženo poznavanju konteksta in ozadja samega besedila, s tovrstno dostopnostjo elektronskih virov le še poudarjeno.