Digital Tools and Electronic Texts

A Methods Network Working Paper

The designation of this document as a ‘working paper’ is an acknowledgement that its content is not meant to be regarded as finalised or fixed. As part of the Methods Network remit to encourage discussion about the advanced use of ICT tools and methods for arts and humanities research, comments, annotations, corrections and recommendations relating to this paper are sought from all those involved with the preparation and use of electronic texts, either in the form of critical editions or in less structured or less mediated formats. The Methods Network has developed a community platform, www.arts-humanities.net to facilitate discussion of this nature and the content of this paper and others in the series will be made available for annotation and comment in a series of wiki articles.

The principle areas that this paper will focus on are the digital tools and techniques that have been developed to acquire, process, analyze and present text in digital formats. For the purposes of this paper, the texts in question are originally from analogue sources and are likely to be works of literature, non-fiction historical documentation (e.g. newspapers, government records), manuscripts, religious writings, etc.

Electronic Publication

As with many activities related to scholarly research, the production of electronic editions and archives - and the associated focus on technologies to assist with that process - has been closely (though not exclusively) entwined with developments associated with the World Wide Web since the mid 1990’s. As the data available to users of the Web has exponentially grown, so has the expectation that material previously only to be found by browsing library stacks should automatically become freely available to all online. In some senses this has actually happened with initiatives such as Project Gutenberg,[1] which provides reading copies of a significant number and range of publications, but it quickly becomes apparent that there is little by way of scholarly apparatus to describe the derivation or the potential inaccuracy of these resources. As such, they are problematic to wholeheartedly endorse as source material on which to base serious and sustained research.

The main alternative to online publishing has been (and in many cases still is) putting material onto a CD (or DVD) but some would argue that this method of delivery creates as well as solves problems. At a recent Methods Network seminar on ‘Text Editing, Scholarship, Books and the Digital World’,[2] one of the conclusions reached by participants was that publishing on CD was expensive and that the media are prone to failure sooner or later, resulting in time-consuming and expensive wrangling between the publisher and the consumer. On the basis that the consumer will want the electronic version to act as a surrogate printed copy, and would naturally expect a printed copy to last a lifetime, any lack of robustness in the electronic version was understood to be very significant in terms of user satisfaction, even after a considerable period of time had elapsed between the purchase and use of the item.

This issue need not, however, be an inhibiting factor to publication, just as issues to do with technological obsolescence of web based materials need not represent insuperable obstacles to the long term viability of online resources. The key to avoiding such issues is by building in sustainability at the outset, so that new releases of the material, either for re-publication onto new more stable media, or for release into new editions of web browsers (where older code is deprecated to the point where it no longer displays effectively) is not only possible but built into the strategic project plan. The first of these contingencies

[1] Project Gutenberg, http://www.gutenberg.org/wiki/Main_Page, (accessed 28 June 2007)
[2] Text Editing, Scholarship, Books and the Digital World, A Methods Network Seminar, King’s College, London, 14 July 2006. Report accessible at: http://www.methodsnetwork.ac.uk/redist/pdf/es3_2rapreport.pdf, (accessed 28 June 2007)

might be accommodated by technical strategies and will be addressed in later sections of this paper. The second, involving 'planning', is rather more difficult to prescribe, relating as it does to more uncertain territories to do with budget availabilities, the effectiveness and audience for the resource, and the perception of the value of the product and its scholarly contribution – none of which will necessarily be predictable at the commencement of a project.

Despite, or perhaps *because* of these difficulties, it may be useful to look at some exemplar projects that have been influential models for electronic publishing and have also helped to shape the use and development of tools for the processing and manipulation of text. Please refer to the appendix section at the end of this paper for some examples of high-quality resources that have been produced over the last decade or so, and which indicate some of the achievements and advances that have been made in the field of digital text editing. A great many other projects could equally feature on this shortlist, including the work being carried out to present the 28th edition of the Nestle-Aland Greek New Testament,[3] an initiative that Peter Robinson describes as 'perhaps the most elaborate and ambitious of all current electronic edition projects'.[4] A few other projects and centres of activity are listed in the 'web resources' section at the end of this paper.

As a framework for looking at this area of activity, it may be useful to consider individual elements of the information cycle as they apply to the task of preparing an electronic edition (see fig. 1), to see where and how different tools, including those normally associated with other disciplines, can support this process.

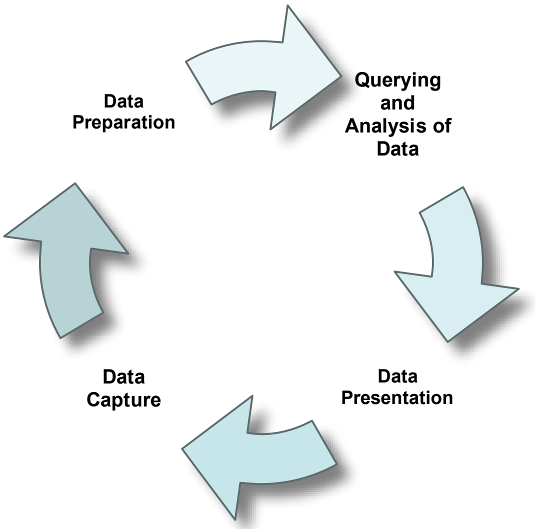


Fig. 1 The data information cycle

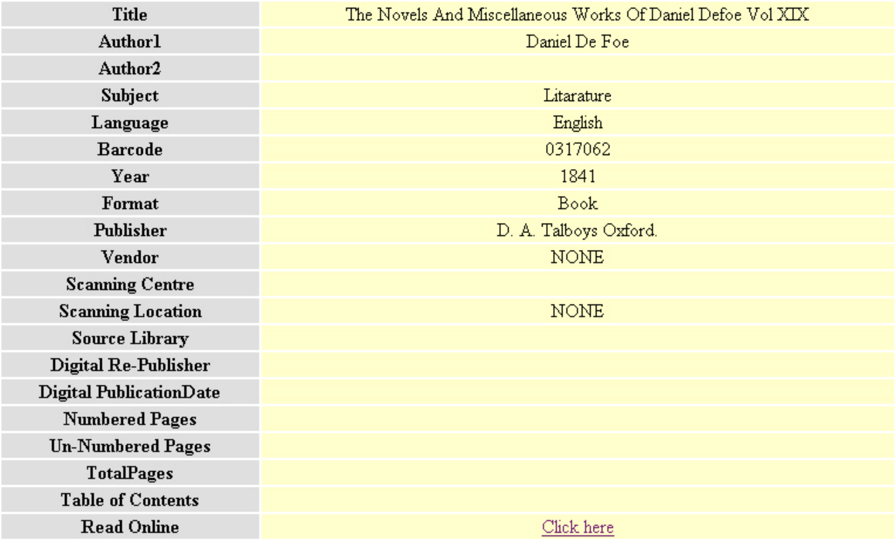
Data Capture

As detailed in one of the other working papers in this series (Tools and Methods for Historical Research[5]) digital methods for the bulk acquisition of data from printed material into digital formats have now been developed to the point where very large scale archives comprising of scores of millions of items are not only a possibility but are becoming a reality. The JISC funded Eighteenth Century Parliamentary Papers

[3] University of Munster Institute for New Testament Textual Research, http://nestlealand.uni-muenster.de/, (accessed 28 June 2007)
[4] Peter Robinson, http://www.digitalmedievalist.org/article.cfm?RecID=6, (accessed 28 June 2007)
[5] Methods Network Working Paper, http://www.methodsnetwork.ac.uk/resources/workingpapers.html, (accessed 28 June 2007)

project based at the University of Southampton[6] has installed scanning equipment that is capable of a throughput of one million items a year and achieves this by using vacuum enabled page turning equipment and laser guided edge detection sensors to identify the borders of pages. In conjunction with automated Optical Character Recognition software (OCR)[7], this project intends to digitise all the original printed parliamentary materials from the eighteenth century and to make them freely available on the Web.

The levels of automation that are being used on digitization projects vary widely according to their respective objectives and there is much debate about what might be considered the most appropriate or effective techniques, both for sustaining large amounts of digital archival material far into the future, and in terms of what the most suitable format for delivery of resources is right now. The aim of an initiative such as the Carnegie Mellon Universal Library[8] is to amass a very substantial number of texts that will act as a globally accessible digital repository of written language. The motivation for doing this is expressed in terms of philanthropy and preservation and as such, does not set out to provide the sort of data about the scanned object that might be useful to someone involved with advanced literary research. A random example of the fullest available catalogue record for a particular item is shown in fig. 2.



Title	The Novels And Miscellaneous Works Of Daniel Defoe Vol XIX
Author1	Daniel De Foe
Author2	
Subject	Litarature
Language	English
Barcode	0317062
Year	1841
Format	Book
Publisher	D. A. Talboys Oxford.
Vendor	NONE
Scanning Centre	
Scanning Location	NONE
Source Library	
Digital Re-Publisher	
Digital PublicationDate	
Numbered Pages	
Un-Numbered Pages	
TotalPages	
Table of Contents	
Read Online	Click here

Fig. 2 A catalogue record from the Carnegie Mellon Universal Library

With the possibility of clicking on a link to access various versions of the text itself, presented either as an image of the original page, or transcribed into plain or HTML text, this offers the reader an extremely useful source of easy reference for a variety of tasks, and generating material for this purpose can be done using ‘rough’ or ‘dirty OCR’ techniques.[9] Used in the Making of America[10] and American Memory[11] projects (for example), the search and retrieval functions are based on unchecked OCR generated text which sits behind images representing the original pages of the object, and for the rapid generation of very large archives with limited searching mechanisms, this is clearly very effective. One obvious drawback is that unchecked automatically generated text is only going to provide approximate information retrieval accuracy.

[6] University of Southampton, 18th Century British Parliamentary Papers, http://www.bopcris.ac.uk/18c/, (accessed 28 June 2007)
[7] The Parliamentary Papers project has selected Abbyy® Fine Reader OCR software, see: http://www.abbyy.com/, (accessed 28 June 2007)
[8] Carnegie Mellon University, Universal Library, http://tera-3.ul.cs.cmu.edu/, (accessed 28 June 2007)
[9] Cited by Willett (2004), as terms coined by John P. Wilkins (p.246)
[10] University of Michigan, The Making of America, http://www.hti.umich.edu/m/moagrp/, (accessed 28 June 2007)
[11] Library of Congress, American Memory, http://memory.loc.gov/ammem/index.html, (accessed 28 June 2007)

An alternative approach is to manually key all relevant text and this is the strategy that has been adopted by the Text Creation Partnership,[12] a collaboration based at the University of Michigan, who are working with images of texts that have been created by three different projects: Early English Books Online (EEBO),[13] Evans Early American Imprint Collection (EVANS),[14] and Eighteenth Century Collections Online (ECCO).[15] The text of the selected material featured from these archives is being carefully encoded using XML tags to define various features such as titles, headings, notes, stage directions, captions, acts and verses etc.

Adopting what might be considered an intermediate approach between these two positions, the Nineteenth Century Serials Edition Project,[16] a collaboration between Birkbeck College, the British Library and King's College London, have teamed up with a commercial company, Olive Software, to implement a system that attempts to automatically apply XML encoding to the text contained within scanned images. The software, ActivePaper Archive™ uses a combination of intelligent segmentation and a whole array of algorithms particularly attuned to accommodate the difficult range of layouts found in historic newspapers to provide a searchable textbase. The poor quality printed texts are subject to enhancement using fuzzy logic and probabilistic matching processes resulting in significantly more accurate results than standard OCR scanning.[17]

Historians refer to the recognition of the importance of the original context of information as a 'source-oriented' approach and it is equally valuable for those involved with literary research to be able to see and analyse the original object in sufficient detail. An adequately high resolution image might, for instance, mean a scholar will be able to understand some of the physical properties of the object such as the type of paper used, the method of binding, or the printing technique employed in the production of a book or manuscript. The standards of digital image capture employed by projects such as the Digital Image Archive of Medieval Music (DIAMM)[18] and the Codices Electronici Sangallenses (CESG) Virtual Library,[19] are both good examples of online resources that enable advanced research through high resolution data capture. In the context of text editing, a highly desirable feature is to be able to link a specific part of the image to the corresponding unit of text in a manually or automatically created transcription. Implementing this for an entire page is relatively trivial, but at the level of a single word, the manual intervention required to identify the precise part of the image corresponding to every word can be very onerous. One potential solution to this problem is to apply OCR in reverse, whereby the software starts with the disambiguated word and attempts to find a pattern on the image file that corresponds with that text.

Data Preparation

Using XML (Extensible Markup Language) encoding to add descriptive value to digitised textual material is perhaps about as commonplace an activity throughout humanities computing as word processing is more generally across all subject disciplines. In the context of a paper on text editing however, it may be worthwhile rehearsing at least a few of the reasons why this practice is so widespread, if for no other reason, so that a foundation can be established for subsequently addressing more applied XML techniques. One particular implementation of XML that has gained a remarkable level of international

[12] University of Michigan, Text Creation Partnership, http://www.lib.umich.edu/tcp/, (accessed 15 January 2007)
[13] Chadwyck-Healy, Early English Books Online, http://eebo.chadwyck.com/home, (accessed 28 June 2007)
[14] University of Michigan, Evans Early American Imprint Collection, http://ets.umdl.umich.edu/e/evans/, (accessed 28 June 2007)
[15] Gale, Eighteenth Century Collections Online, http://www.gale.com/EighteenthCentury/, (accessed 28 June 2007)
[16] Nineteenth Century Serials Edition Project, http://www.ncse.kcl.ac.uk/index.html, (accessed 15 January 2007)
[17] A detailed project report and a description of the software is available at: Olive Software, http://www.uk.olivesoftware.com/conference/Project_Report_A4.pdf, (accessed 15 January 2007)
[18] Digital Image Archive of Medieval Music, http://www.diamm.ac.uk/, (accessed 15 January 2007)
[19] Universitas Frieburgensis, Codices Electronici Sangallenses, http://www.cesg.unifr.ch/en/index.htm, (accessed 28 June 2007)

acceptance is the Text Encoding Initiative (TEI), a set of guidelines originally developed using the SGML (Standardised General Markup Language) standard but updated to be compatible with XML when it replaced its predecessor as the widely used markup scheme of choice. The TEI guidelines allow those working with texts wide latitude in terms of how lax or precise they wish to be in describing their chosen resource, but the best returns on effort expended are generally had by using as much of the intrinsic detail of the system as possible and by applying it rigorously throughout the data. Jerome McGann in his influential essay, 'The Rationale of Hypertext'[20] argues vividly in favour of the application of electronic methods to literary studies and states some of the advantages to be gained from their application, as opposed to more traditional print-based techniques

- Internal links and data relationships within the resource can be minutely and accurately defined
- Previously distributed elements (unavoidable in printed formats of the work) can be made simultaneously present to each other
- A larger quantity of material can be addressed and complex navigational routes through this mass allowed
- Open ended and collaborative editions of works can be envisioned rather than having to commit to laborious and expensive reprinting of paper based copies

Whilst his remarks embrace tools generally, the 'cement' that binds the information together and enables linking elements to be embedded into information is most often some implementation of XML (often incorporating TEI compliance) and is thus central to any strategy that the supporters of digital editions will promote to increase the perception of these resources as the most logical and sensible way to compile this type of complex scholarly work.[21]

Whilst the XML/TEI model is pervasive and powerful in the way that it can be implemented, the intrinsic complexity of written data, particularly poetry and certain forms of prose literature, require encoding models that include a semantic and syntactic flexibility that regular implementations of markup language can struggle to accommodate, based as they are primarily on tree-like structures and strictly nested components. The widely acknowledged problem of accommodating overlapping hierarchies in XML/TEI has produced a number of potential solutions and is the focus of a TEI Special Interest Group (SIG) who maintain a wiki listing a number of systems which provide examples of extended or alternative models.[22]

MECS (Multi-element Coding System) is one of these alternative schemes, proposed by Michael Sperberg-McQueen (incidentally one of the principal architects of the TEI) and Claus Huitfeldt, who developed it for the Wittgenstein archives in Bergen, Norway, in response to the complex multi-hierarchical nature of that dataset. This system relies on a slightly different tag format which enables non-nested passages to be marked-up. A further development of this system is TexMECS where the encoding defines the following features:

1) Empty elements marked by sole-tags
2) Normal elements with start and end tags
3) Interrupted elements with start, suspend, resume and end tags
4) Elements with children whose order has no significance and which can therefore be reordered
5) Virtual elements, which have a generic identifier and attributes, and who share children with another element in the document
6) Self-overlapping elements, which use a simple co-indexing scheme: tags are co-indexed by a tilde and a suffix of numbers and letters[23]

[20] McGann (2001)
[21] For a commentary on the status of digital editions in 2005, see: Peter Robinson, http://www.digitalmedievalist.org/article.cfm?RecID=6, (accessed 15 January 2007)
[22] Text Encoding Initiative Wiki, http://www.tei-c.org.uk/wiki/index.php/SIG:Overlap, (accessed 28 June 2007)
[23] Cover Pages, Markup Language for Complex Documents (Bergen MLCD Project), http://xml.coverpages.org/mlcd.html, (accessed 28 June 2007)

Every TexMECS document should be translatable into a GODDAG (General Ordered-Descendant Directed Acyclic Graph)[24] structure which is another proposal by Sperberg-McQueen and Huitfeldt which makes use of graph theory to navigate around the problem of overlap and provide individual entities with additional nodes as points of relation within a tree-like structure. The graph resembles a tree, but differs from it in that multiple parent nodes can contain the same child.

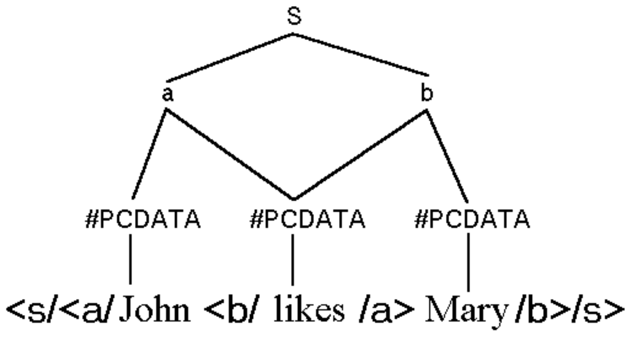
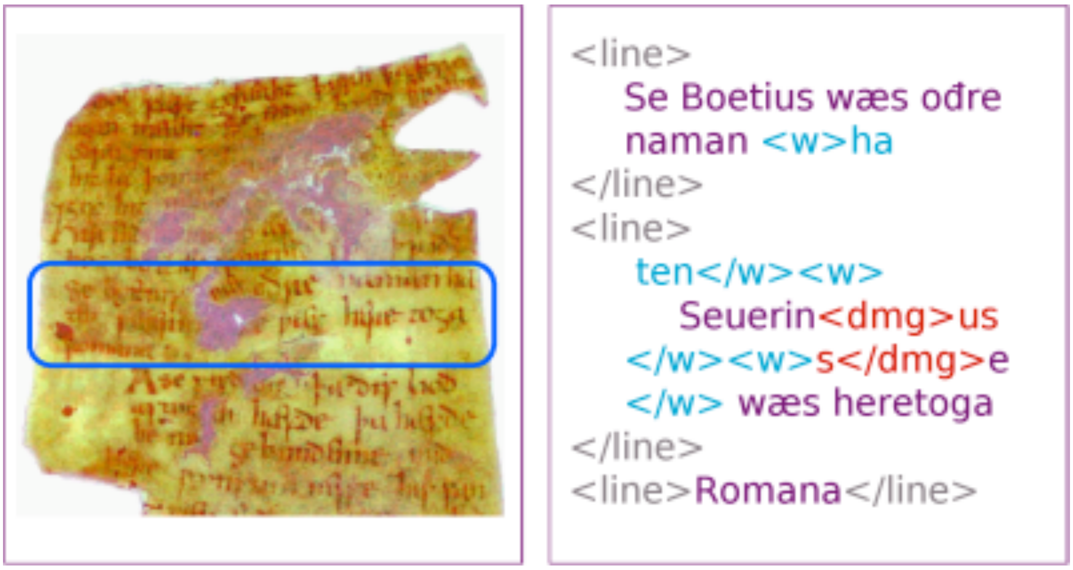


Fig. 3 Sample representation of a GODDAG structure[25]

The ARCHway project,[26] based at the University of Kentucky, has also looked at the problem of overlapping markup and has developed a package called the Edition Production Technology (EPT) to deal with the type of tagging problems associated with the image based resources that the project focuses on.



```
<line>
    Se Boetius wæs odre
    naman <w>ha
</line>
<line>
     ten</w><w>
        Seuerin<dmg>us
    </w><w>s</dmg>e
    </w> wæs heretoga
</line>
<line>Romana</line>
```

Fig. 4 Illustration of the use overlapping markup[27]

[24] World Wide Web Consortium, GODDAG: A Data Structure for Overlapping Hierarchies, http://www.w3.org/People/cmsmcq/2000/poddp2000.html, (accessed 28 June 2007)
[25] Claus Huitfeldt and Michael Sperberg-McQueen, http://www.iath.virginia.edu/ach-allc.99/proceedings/sperberg-mcqueen.html, (accessed 28 June 2007)
[26] University of Kentucky, The Archway Project, http://beowulf.engl.uky.edu/~kiernan/ARCHway/entrance.htm, (accessed 28 June 2007)
[27] Alex Dekhtyar et al, Building Image-based Electronic Editions using the Edition Production Technology, http://www.ieee-tcdl.org/Bulletin/v2n1/dekhtyar/dekhtyar.html, (accessed 28 June 2007)

The method implemented in EPT is called Concurrent Markup Hierarchies (CMH) which manages the encoding through Extended XPath, an extension of regular XPath, which also takes advantage of the node method defined by the GODDAG structure. XPath[28] is a W3C recommended method of modelling XML documents as trees of nodes, such that individual parts of the document can effectively be addressed, which in turn allows other processes such as XSLT (Extensible Stylesheet Language Template) to then carry out transformations on the data contained at that node. Another notable feature of EPT is that it is based on an Eclipse platform that features a plug-in architecture offering extension points for new code to be added which will extend the existing functionality and present opportunities for collaborative work at a range of different levels.

The difficulty of presenting effectively marked up text reflects the intellectual challenge of analysing the source material. One of the most useful and positive scholarly aspects of applying markup to a humanities text is simply to gain a better understanding of all of its complexities and its structure at a very fine level of detail. Another major challenge specifically faced by those preparing digital editions is how to collate the variations of any given text so that a detailed picture of all of the deviations between those different versions becomes apparent. Although this sounds simple in principle, actually carrying out the task involves considerable work, particularly where there are numerous variants to juggle and where it is difficult to precisely locate and correlate the amendments from one manuscript to another. It was as a response to this problem that Peter Robinson first developed the COLLATE[29] program in 1989 when he was working on an edition of two ancient Norse poems which existed in forty four separate manuscripts. He devised a Macintosh font to transcribe the characters of these manuscripts and found it easier to edit the closest version rather than begin again from scratch with the next version. This principle is akin to editing from a ‘base text’ and informed the subsequent design of the system.[30] The current version of the programme can collate 2000 simultaneous variant texts and, is informing the design of EDITION, the next generation tool that Robinson is developing.[31]

IATH (Institute of Advanced Technology in the Humanities) at the University of Virginia have recently announced a new tool called JUXTA[32] which also carries out collation tasks and was developed principally with nineteenth and twentieth century textual material in mind. As well as providing similar functions to COLLATE in that variant witnesses can be compared against a base-text which can be replaced with an alternative and recompiled at any time, it also features analytical visualization tools that include a ‘heat’ map of all textual variants and a histogram of collations. The latter displays the density of all variation from the base text which is particularly useful for long texts.

Whether the above tools are in fact ‘data preparation’ or ‘query and analysis’ tools is a moot point, but to finalise this section, it will be worth mentioning one more influential tool that has a collation module amongst its range of functions. TuStep (TUebingen System of Text Processing Programs) was developed by Wilhelm Ott at the University of Tuebingen in the late 1960’s and has been used in a huge number of research projects. Its longevity bears witness to an ongoing need for the functionality it delivers as well as the development work that has clearly been put into the system to keep it relevant to communities of users engaging with SGML and XML.

Querying and Analysis of Data

[28] World Wide Web Consortium, XPath, http://www.w3.org/TR/xpath, (accessed 15 January 2007)
[29] University of Birmingham, COLLATE, http://www.itsee.bham.ac.uk/software/collate/, (accessed 15 January 2007)
[30] A very useful description of the programme and its application is available at: University of Virginia Library Electronic Text Centre, http://etext.virginia.edu/services/helpsheets/software/collate2.html, (accessed 15 January 2007)
[31] Scholarly Digital Editions, http://www.sd-editions.com/EDITION/, (accessed 28 June 2007)
[32] University of Virginia, Juxta, http://www.patacriticism.org/juxta/, (accessed 28 June 2007)

The range of functionality that TuStep offers makes it the ideal candidate to begin a section devoted to the range of tools and techniques that are (or will be) available to scholars working with digital texts. In addition to the collation function referred to above, the other categories of operation are:

- Editing
- Processing Text
- Preparing Indexes
- Presorting
- Sorting
- Generating Indexes and Concordances
- Generating Listings
- Typesetting[33]

With additional file handling and job control commands, TuStep provides a modular system where programmes are run in sequence to cover the whole chain of processes, whilst enabling the editor to intervene at any stage. Batch processes, defined and controlled by user input parameters can be aggregated and then stored so that the same sequence can be used subsequently. In his discussion of 'Text Tools', John Bradley attempts to be realistic when summarising the level of expertise required to use this tool.

> Tustep is especially developed for performing tasks of interest to scholars working with texts, although it will still take real work to learn how to use it. […] For non-programmers, then, both TuStep and Perl have steep and long learning curves.[34]

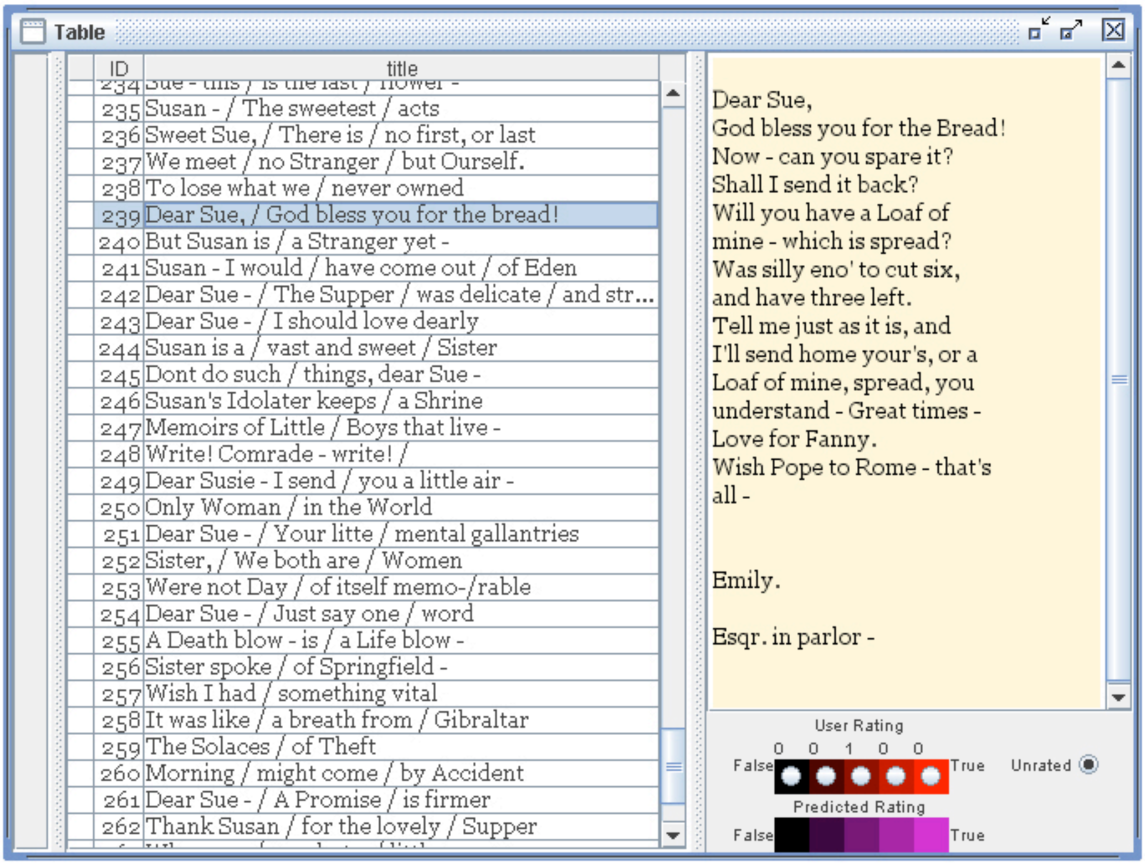
His mention of Perl is significant in terms of text processing as it is often cited as the programming language that is particularly suited to the manipulation of text files, having as it does, a wealth of pattern-matching and string-handling constructs which are of practical use for editorial and linguistics activities.

The ongoing development of EDITION[35] by Peter Robinson and colleagues (software based partly on the COLLATE system as stated above) will offer researchers a similarly full-featured toolset for the production of digital editions. The objective of the project is to design a system that will be usable by any scholar who has the knowledge to produce a print edition whilst still featuring enough functionality to enable output of exemplary quality. The other components of the system will be based on ANASTASIA,[36] (also developed by Robinson) which is designed for publishing large and highly complex XML documents; and a third piece of software developed by the ARCHway[37] project, based at the University of Kentucky, that allows users to link text and images down to the finest level of detail.

A feature of Robinson's research in the past has been the use of cladistic analysis on manuscripts. Borrowed originally from evolutionary biology, this method attempts to map 'trees of descent or history for which the fewest changes are required, basing this on comparisons between the descendants'.[38] Using PAUP (Phylogenetic Analysis Using Parsimony) software, family trees of words were created showing patterns of similarity and deviation in regularized word lists.[39] Hockey also reports on the use, by Patricia

[33] Definitions of these functions are available at: University of Tuebingen, Textdata Processing with TUSTEP, http://www.zdv.uni-tuebingen.de/tustep/tdv_eng.html, (accessed 28 June 2007)
[34] Bradley (2004) pg. 521
[35] Scholarly Digital Editions, http://www.sd-editions.com/EDITION/, (accessed 15 January 2007)
[36] Anastasia, http://anastasia.sourceforge.net/index.html, (accessed 15 January 2007)
[37] University of Kentucky, Archway Project, http://beowulf.engl.uky.edu/~kiernan/ARCHway/entrance.htm, (accessed 28 June 2007)
[38] Hockey (2000) pg.130
[39] This is referred to in a case study available at: AHDS, Recounting Digital Tales: Chaucer Scholarship and The Canterbury Tales Project, http://ahds.ac.uk/creating/case-studies/canterbury/index.htm, (accessed 28 June 2007)

Galloway (1979),[40] of cluster analysis and dendrogram diagrams; techniques which appear to have been transposed from her activities relating to archaeology, a discipline in which cluster analysis is a standard technique.



ID	title
234	Sue - this / is the last / flower -
235	Susan - / The sweetest / acts
236	Sweet Sue, / There is / no first, or last
237	We meet / no Stranger / but Ourself.
238	To lose what we / never owned
239	Dear Sue, / God bless you for the bread!
240	But Susan is / a Stranger yet -
241	Susan - I would / have come out / of Eden
242	Dear Sue - / The Supper / was delicate / and str...
243	Dear Sue - / I should love dearly
244	Susan is a / vast and sweet / Sister
245	Dont do such / things, dear Sue -
246	Susan's Idolater keeps / a Shrine
247	Memoirs of Little / Boys that live -
248	Write! Comrade - write! /
249	Dear Susie - I send / you a little air -
250	Only Woman / in the World
251	Dear Sue - / Your litte / mental gallantries
252	Sister, / We both are / Women
253	Were not Day / of itself memo-/rable
254	Dear Sue - / Just say one / word
255	A Death blow - is / a Life blow -
256	Sister spoke / of Springfield -
257	Wish I had / something vital
258	It was like / a breath from / Gibraltar
259	The Solaces / of Theft
260	Morning / might come / by Accident
261	Dear Sue - / A Promise / is firmer
262	Thank Susan / for the lovely / Supper

Dear Sue,
God bless you for the Bread!
Now - can you spare it?
Shall I send it back?
Will you have a Loaf of
mine - which is spread?
Was silly eno' to cut six,
and have three left.
Tell me just as it is, and
I'll send home your's, or a
Loaf of mine, spread, you
understand - Great times -
Love for Fanny.
Wish Pope to Rome - that's
all -

Emily.

Esqr. in parlor -

Fig. 3 Screenshot from the NORA Project system

The widespread use of data (or text) mining techniques in humanities projects is now fairly well established and its application to literary studies can be demonstrated by reference to the NORA project,[41] a collaborative U.S. based project directed by John Unsworth. The objective is to 'to produce software for discovering, visualizing, and exploring significant patterns across large collections of full-text humanities resources in existing digital libraries.'[42] Fig. 3 shows a screenshot from one of the functions that the project is developing which allows users to put training information into the system that rates the text for evidence of a specific attribute. Once a training set is established, the system will then search all of the available data and apply the knowledge that has resulted from an analysis of the training set to return relevant results for data the user has not evaluated. The principles of data mining have emerged from computing science and encompass a variety of complex methods and procedures, but at a very general level, it is clear that intelligent knowledge-augmented searching is already being used in a variety of disciplines and may well provide new methods of querying the ever larger datasets that arts and humanities scholars are confronted with.

[40] Galloway bibliography at: University of Texas, School of Information, http://www.ischool.utexas.edu/~galloway/biblio.html, (accessed 28 June 2007)

[41] NORA Project, http://www.noraproject.org/, (accessed 15 January 2007), see also the Methods Network Working Paper on 'Tools for Historical Research' which features a section on 'Data Mining'

[42] NORA Project, http://www.noraproject.org/description.php, (accessed 28 June 2007)

The discipline of linguistics is one such area that grapples with increasingly larger repositories of information, frequently in the form of corpora that in some cases contain several hundred million words.[43] Text editing and literary scholarship can clearly benefit from these vast reference sources, particularly in relation to historical collections of words which have often been harvested from literary sources.[44] A recent Methods Network workshop on *Corpus Approaches to Literature* demonstrated a range of techniques that will be of interest to literary scholars. Focused on the use of Wordsmith,[45] a lexical analysis software package developed by Mike Smith at the University of Liverpool, participants were introduced to clustering, collocation, colligation and semantic prosody analysis methods that can shed light on a number of issues including style and content analysis, attribution, the study of literary effects and the creative use of language in comparison with quantitative norms.[46] In one study concerning the use of repetition in the works of Charles Dickens, Wordsmith was used to pick out short phrases that featured repeatedly in: a single chapter of a Dickens novel; the whole book; and then in a number of other novels as featured in a corpus of nineteenth century literature. It was demonstrated that Dickens recycled phrases more regularly than any of the other authors featured in the corpus, which might represent quantitative proof of the effect that journalistic deadlines had on Dickens's often serialised output

In the course of other Methods Network events,[47] presentations have been given on a range of techniques that also have relevance. The use of the automatic tagging system CLAWS[48] (the Constituent Likelihood Automatic Word-tagging System) is widely referenced and was the system used to automatically add part of speech (POS) tags to the British National Corpus (BNC). The variant spelling detection programme VARD [49] (Variant Detector) uses fuzzy matching procedures to try and identify and match historical spellings of words with their 'normalized' equivalents. The Historical Thesaurus of English[50] provides researchers with an enormous resource, arranged semantically and chronologically, that details English vocabulary as it has changed over the centuries. The use of a semantic ontology, USAS (UCREL Semantic Analysis System) has also been featured in research connected with word domain analysis in Shakespeare.[51] This work concentrates on the semantic tagging of texts which allows for the grouping of words into conceptual clusters.

Data Presentation

Specific tools for manipulating data into formats for publication have already been mentioned with reference to ANASTASIA, TuStep and EDITION. The first of these is a currently available dedicated tool and labours under the full title, Analytical System Tools and SGML/XML Integration Applications. Where it differs from other SGML/XML publishing systems is that in addition to recognising the primary document hierarchy as expressed by the encoding, it also is capable of 'reading' the text according to its left to right relation in the document stream, by column, by page or indeed from any point in the text to any other point. This gets around the problem of not being able to take into account multiply hierarchical or overlapping

[43] The Bank of English or COBUILD corpora was reported as containing 524 million words in 2005
[44] For a list of corpora see: David Lee, http://devoted.to/corpora, (accessed 28 June 2007)
[45] Mike Scott, Wordsmith, http://www.lexically.net/wordsmith/, (accessed 15 January 2007)
[46] For a report on this workshop see: Methods Network, http://www.methodsnetwork.ac.uk/redist/pdf/act3report.pdf, (accessed 28 June 2007)
[47] Methods Network, *Word Frequency and Keyword Extraction*, An Expert Seminar on Linguistics http://www.methodsnetwork.ac.uk/activities/es01mainpage.html, (accessed 28 June 2007)
And *Historical Text Mining*, a Methods Network Workshop
http://www.methodsnetwork.ac.uk/activities/act6.html
[48] Lancaster University, CLAWS, http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/, (accessed 28 June 2007)
[49] Lancaster University, Workshop on Historical Text Mining, http://ucrel.lancs.ac.uk/events/htm06/, (accessed 28 June 2007)
[50] University of Glasgow, Historical Thesaurus of English, http://www.arts.gla.ac.uk/sesll/englang/thesaur/thes.htm, (accessed 28 June 2007)
[51] Methods Network, http://www.methodsnetwork.ac.uk/redist/pdf/es1_08archer.pdf, (accessed 28 June 2007)

sections of text (which SGML/XML/TEI struggles to accommodate – see above). On the ANASTASIA sourceforge web page,[52] it states that the programme is 'an event-driven procedural environment for handling XML document collections' and also gives a link to a page that highlights the difference between it and XSLT (Extensible Stylesheet Language Template), which at a casual glance appears to fulfil the same function.

The use of ANASTASIA is recommended for large collections of data with complex structures where there is a requirement to build real-time views of information that is scattered throughout that data, and which cuts across XML element descriptions. For alternative scenarios where the hierarchical structure of the XML document presents no such difficulties, the most widely referenced XSLT processing systems appear to be SAXON[53] and XALAN.[54] These use the XSLT transformation vocabulary, in association with XPath (a language for addressing parts of XML documents), to rearrange, sort, combine and transform one XML document into another; or alternatively output that content into HTML or text file formats. XSLT can also be used in conjunction with Cascading Style Sheets (CSS), a W3C[55] recommendation since 1996 which concentrates on the specification of the format of HTML (and SGML/XML) documents as they appear in a browser, but which lacks the sophisticated functionality of XSLT/XPath techniques to query and manipulate data.

The Versioning Machine[56], conceived by Susan Schreibman and based at the University of Maryland, is a web-based system for displaying and comparing different version of the same text. It supports the display of XML texts encoded according to the guidelines of the TEI and allows different witness texts to be displayed side-by-side (a diplomatic edition next to a manipulable image of the witness text for example) in addition to features such as an enhanced typology of notes, synchronized scrolling and line matching functions. The system will accommodate separately encoded TEI documents of poems and prose and will display these files side-by-side but will be unable to take advantage of the synchronised scrolling and line matching features. Alternatively, one can use the TEI's "critical apparatus tagset" (TEI.textcrit) to encode all the witnesses in one XML file, thereby cutting down on the amount of repetitive encoding required and enabling all the functionality of the Versioning Machine, but at the cost of added complexity at the initial encoding stage.

At a more general level it may be worth briefly focusing on the use of Fedora[57] (Flexible Extensible Digital Object and Repository Architecture) for the management and delivery of a wide range of digital content, particularly multiple versions of objects. In use by a number of organisations including Tufts University (hosting the Perseus Digital Library[58]) and the University of Hull, it is an open source institution-level repository system that provides a digital asset management architecture. It provides a framework for a variety of resources encompassing digital archives and multimedia authoring systems, an example of the latter being the DPubs system developed by the Cornell University Library (a full-featured extensible publishing platform for the organisation, presentation and delivery of scholarly material).[59]

Conclusion

[52] ANASTASIA, http://anastasia.sourceforge.net/anaprocess.html, (accessed 15 January 2007)
[53] SAXON, The XSLT and XQuery Processor, http://saxon.sourceforge.net/, (accessed 28 June 2007)
[54] The Apache XALAN Project, http://xalan.apache.org/, (accessed 28 June 2007)
[55] World Wide Web Consortium, http://www.w3.org/, (accessed 28 June 2007)
[56] University of Maryland, The Versioning Machine, http://v-machine.org/index.php, (accessed 28 June 2007)
[57] Fedora Project, http://www.fedora.info/, (accessed 28 June 2007)
[58] Tufts University, Perseus Project, http://www.perseus.tufts.edu/, (accessed 28 June 2007)
[59] David Ruddy, The Fedora Users conference at the University of Virginia, 2006, http://www.lib.virginia.edu/digital/fedoraconf/abstracts.shtml, (accessed 28 June 2007)

The emergence of Web 2.0 applications has focused attention on an area that might be defined as a many-to-many publishing model, which includes such activities as collaborative tagging, shared resource building, folksonomies, social networking, social bookmarking, collaborative editing (wikis), podcasting and RSS feed delivery. As with all other areas of arts and humanities research, the impact that this new breed of techniques may have on scholarship is, as yet, uncertain in its specificity but almost inevitable in its generality. Peter Shillingsburg has put forward a proposal for what he refers to as a 'Collaborative Literary Research Electronic Environment', which would operate as a 'knowledge site' consisting of interconnecting modules offering researchers the chance to create dynamic and interactive works in a shared scholarly space.

Peter Robinson has also expressed a view on the direction scholarly editions might take in the future. He refers to:

> …the making of what may be called fluid, co-operative and distributed editions. These editions will not be made or maintained by one person or by one group, but by a community of scholars and readers working together: they will be the work of many and the property of all.[60]

As should be apparent from the preceding sections, the enthusiasm and ingenuity of practitioners in the field of digital text editing - in association with colleagues from other disciplines - has led to the development of a useful and interesting range of tools and methods for the preparation, analysis and presentation of textual data in scholarly formats. It is, however, difficult to ignore the fact that some of these practitioners have professed dissatisfaction at the levels of interest shown in digital editions by the wider community of literary scholars.

At a recent Methods Network workshop[61], Sharon Ragaz in her rapporteur's summary concluded that from a number of remarks made by various speakers throughout the day, there was clearly a gap between the producers and the users of these resources, the inference being that messages about the utility and the quality of scholarly digital editions are generally not reaching the audience for which they are intended – or if they are, are falling on deaf ears because of a residual preference for printed formats, based perhaps, on aesthetic and ergonomic factors rather than issues to do with function and facility. Edward Van Houtte illustrated this by pointing out that whilst electronic editions were of 'inestimable value' as academic resources, as cultural product they were valueless. Whether true or not, it effectively makes the point that the principle purpose of digital editions is to take full advantage of the extra analytical value that can be extracted from the presentation of material in this format, and that they should be seen as a necessary tool to sit alongside the precious first edition printed copy of a text, rather than an either/or decision.

The Model Editions Partnership[62] was a consortium of twelve editorial projects publishing a range of historical documentation that was designed not only to make data available to scholars, students and the public, but also to pursue an agenda that promoted the production and dissemination of electronic editions and prescribed methods by which this could be achieved. The legacy of this project is perhaps uncertain, but as well as providing useful reference material, it also sets an interesting precedent upon which similar initiatives to raise the profile of electronic editions could be based.

Neil Grindley
Senior Project Officer
Methods Network   Version control – 29 June 2007

[60] Peter Robinson, Where we are with electronic scholarly editions, and where we want to be, http://computerphilologie.uni-muenchen.de/jg03/robinson.html, (accessed 28 June 2007)
[61] Methods Network, Expert Seminar on Literature, Text Editing in a Digital Environment, Centre for Computing in the Humanities, King's College, London, 24 March 2006
[62] Model Editions Partnership, http://adh.sc.edu/, (accessed 28 June 2007)

APPENDIX

The Rossetti Archive

http://www.rossettiarchive.org/index.html

Directed by Jerome McGann at the University of Virginia, this archive aims, by 2008, to publish all text and images produced by Dante Gabriel Rossetti (1828 – 1882) along with a contextual corpus of material relating to his work, taken from contemporaneous and historical sources.[63] Currently in its third phase of development, this project is a useful example of a resource that is adapting to new technologies as they emerge, the latest innovation being the incorporation of a new search engine, based on the open source Lucene platform (a high performance full-featured search engine library written entirely in Java) which also takes fuller advantage of the fact that the projects source files have recently been converted to XML formats. The home page states that 'a full scale re-design of the Rossetti Archive is currently underway'[64] and also reports collaboration with the NINES project,[65] which provides a networked interface for nineteenth century scholarship; so it is clear that ongoing engagement with technology in order to provide a sustainable resource is a clear priority in this project, as is a commitment to developing and using new tools e.g. Collex,[66] (a tool for content aggregation, faceted browsing, folksonomy, and interpretive re-use of digital objects).

The Canterbury Tales Project

http://www.canterburytalesproject.org/

Begun in the late 1980's by Peter Robinson (currently co-director of the Institute for Textual Scholarship and Electronic Editing based at the University of Birmingham), the Canterbury Tales project has so far produced seven CD-ROM publications and has more in the pipeline, countering the charge that publication onto this type of media is overwhelmingly problematic. The collation of the various 'witnesses' (i.e. the variant sources discovered for a specific text) have all been accomplished by the use of a programme that Robinson designed specifically for the task, and which has been used and widely referenced by other editors of scholarly material. Being one of the few tools available for this type of work, COLLATE[67] represents a significant contribution to the discipline and is now acting as a component part of a more comprehensive system that Robinson and others are creating called EDITION.[68] This will incorporate collation as just one function among several that will address the entire production process of scholarly editions.

The Electronic Beowulf Project

http://www.uky.edu/~kiernan/eBeowulf/guide.htm

Building on earlier research into using fibre-optic light to determine illegible fragments of the Beowulf manuscript (which was damaged by fire in 1731), Kevin Kiernan from the University of Kentucky, in collaboration with staff based at the British Museum, embarked on a project to digitise the fragile manuscript, in order to preserve and increase potential access to a unique and valuable document. In addition to the high-quality images of the manuscript itself there are also images from an eighteenth century transcription known as Cotton Vitellius A. xv;[69] a copy of the 1815 first edition with early nineteenth century collations; a comprehensive glossarial index; and a new edition and transcript featuring search facilities which allow interrogation of the underlying SGML marked up text. As an example of an image-

[63] Institute for Advanced Technology in the Humanities, Rossetti Archive, http://www.rossettiarchive.org/index.html, (accessed 28 June 2007)
[64] Institute for Advanced Technology in the Humanities, Rossetti Archive, http://www.rossettiarchive.org/about/index.html, (accessed 28 June 2007)
[65] Networked Infrastructure for Nineteenth Century Electronic Scholarship, http://www.nines.org/, (accessed 28 June 2007)
[66] University of Virginia, http://www.patacriticism.org/collex/, (accessed 28 June 2007)
[67] University of Birmingham, http://www.itsee.bham.ac.uk/software/collate/, (accessed 28 June 2007)
[68] Scholarly Digital Editions, http://www.sd-editions.com/EDITION/, (accessed 28 June 2007)
[69] University of Kentucky, http://www.uky.edu/~kiernan/eBeowulf/descript.htm, (accessed 28 June 2007)

based initiative to represent textual information, this project demonstrates a highly sophisticated cross-disciplinary approach and has informed other initiatives such as the ARCHway Project, 'a universal software platform for [the] creation and maintenance of Image-based Electronic Editions (IBEE)',[70] also based at Kentucky University, which focuses heavily on the use of XML.

The Blake Archive
http://www.blakearchive.org/blake/
Winning prizes under the heading of 'scholarly edition'[71] – although describing itself as a 'hypermedia archive' – this site (based again at the Institute for Advanced Technology in the Humanities at the University of Virginia) uses a range of techniques and technologies which are designed to ensure the longevity of the resource and to maximise its usability and effectiveness. In common with other resources cited, the content is a mixture of images and text and therefore requires a range of strategies to ensure adequate coverage of the material. The technical summary [72] includes references to colour calibration and correction, scanning resolutions, metadata recommendations, and all the usual paraphernalia of image digitisation projects, but also refers to the use of XML and the elaboration of two purpose-designed document type definitions (DTD's) to provide a system of contexts and constraints for all the text information. Additionally, the technical summary describes the use of:

- Apache Cocoon - to facilitate communication between the various functions of the archive
- eXist – an open source native XML database to store, index and search (using xQuery) the XML encoded documents
- Extensible Stylesheet Language (XSL) – to transform the XML documents 'on-the-fly' into HTML pages for display on the Web
- Java applets – to enable two bespoke screen functions (Image Sizer and INote).

References

Printed Sources

Bradley, J., Text Tools, in Schreibman, S., Siemens, R., Unsworth, J., (eds), A Companion to Digital Humanities, (pp.505 - 522)

Hockey, S., Electronic Texts in the Humanities, OUP: New York, 2000

McGann, J., Marking Texts of Many Dimensions, in Schreibman, S., Siemens, R., Unsworth, J., (eds), A companion to Digital Humanities, (pp.198 - 217)

McGann, J., Radiant Textuality,: Literature after the world wide web, Palgrave: New York and Basingstoke, 2001

Nell Smith, M., Electronic Scholarly Editing, in Schreibman, S., Siemens, R., Unsworth, J., (eds), A companion to Digital Humanities, (pp. 306 - 322)

Willett, P., Electronic Texts: Audiences and Purposes, in Schreibman, S., Siemens, R., Unsworth, J., (eds), A companion to Digital Humanities, (pp.240 - 253)

[70] University of Kentucky, http://beowulf.engl.uky.edu/~kiernan/ARCHway/entrance.htm, (accessed 28 June 2007)
[71] The Blake Archive was awarded the 2003 Modern Language Association of America (MLA) prize for 'Distinguished Scholarly Edition'. See: The Blake Archive, http://www.blakearchive.org/blake/MLA.html, (accessed 28 June 2007)
[72] The Blake Archive, http://www.blakearchive.org/blake/public/about/tech/index.html, (accessed 28 June 2007)

Web Resources

Please note: The following web pages are a selection of those consulted during the course of researching this paper. They are cited here as a context for the paper rather than as an attempt to comprehensively map the field. They are sorted by section but are otherwise in no particular order.

Table of Contents

Listings

Scholarly Electronic Publishing Bibliography
http://sepb.digital-scholarship.org/
2,830 articles, books, and other printed and electronic sources that are useful in understanding scholarly electronic publishing efforts on the Internet.

Web 2.0 listing site
http://web2magazine.blogspot.com/2007/01/thanks-for-web-2.html
100 of the best web 2.0 sites

List of Digital Curation Tools
http://www.dcc.ac.uk/tools/digital-curation-tools/
Copious references to tools under a number of useful categories

Textual Scholarship website
http://www.textualscholarship.org/links.html
Links to useful resources

Humanities Computing
http://www.allc.org/imhc/
List of organisations, departments and initiatives

Organisations

AHDS Literature, Languages and Linguistics
http://ahds.ac.uk/litlangling/
Arts and Humanities Data Service

TEI Consortium
http://www.tei-c.org/
Text Encoding Initiative

International Digital Publishing Forum
http://www.idpf.org/
Formerly the Open e-Book Forum

Centre for Textual Scholarship

http://www.cts.dmu.ac.uk/
Based at De Montfort University directed by Peter Shillingsburg

Institute for textual Scholarship and Electronic Editing
http://www.itsee.bham.ac.uk/index.htm
Based at Birmingham University headed by Peter Robinson

Scholarly Digital Editions
http://www.sd-editions.com/
Publisher of high quality digital editions

Society for Textual Scholarship
http://www.textual.org/
Based at MITH (Maryland Institute for technology in the Humanities)

Oxford Text Archive
http://ota.ox.ac.uk/
Searchable and catalogued text repository

Applied Research in Patacriticism
http://www.patacriticism.org/home.html
Jerome McGann and colleagues at the University of Virginia

Digital Medievalist
http://www.digitalmedievalist.org/
International web based community hosted at the University of Lethbridge

Digital Humanities
http://digitalhumanities.org/
Alliance of Digital Humanities Organizations

Projects

BECHAMEL Project
http://eprg.isrl.uiuc.edu/projects.html
Electronic Publishing Research Group page

HyperNietzsche Project
http://www.hypernietzsche.org
To make research freely available on the web

Emily Dickinson Open Review Page
http://www.emilydickinson.org/review/deareview.php
Martha Nell Smith's page on Emily Dickinson

LEADERS Project
http://www.ucl.ac.uk/leaders-project/
Linking Encoded Archival description to electronically retrievable sources

Project Gutenberg
http://www.gutenberg.org/wiki/Main_Page
Volunteer provided texts for books out of copyright

Women Writers Online
http://www.wwp.brown.edu/
Encoded text available from women writers

Model Editions Partnership
http://adh.sc.edu/
Historical digital editions

Making of America
http://www.hti.umich.edu/m/moagrp/
Digitised collection of OCR texts and maps with search function by keyword

Thesaurus Linguae Graecae
http://www.tlg.uci.edu/
All Greek texts from Homer onwards

Internet Shakespeare
http://ise.uvic.ca/index.html
Online edition designed to take advantage of the Internet

The NINES project
http://www.nines.org/
A networked interface for nineteenth century electronic scholarship

The Archway Project
http://beowulf.engl.uky.edu/~kiernan/ARCHway/entrance.htm
System for building libraries of image based scholarly editions

Digtial Nestle-Aland
http://nestlealand.uni-muenster.de/
The forthcoming scholarly edition of the Greek New Testament

The Blake Archive
http://www.blakearchive.org/blake/
Hypertext award-winning archive of Blake's text and images

Universal Library
http://tera-3.ul.cs.cmu.edu/
Carnegie Mellon One Million Book project

Reports & Publications

Electronic Publishing Research Group
http://eprg.isrl.uiuc.edu/publications.html#renear02:DOCENG
Publications from the EPRG at the University of Illinois

Olive Software Conference Page
http://www.uk.olivesoftware.com/conference/
One day conference on newspaper scanning project

Olive Software Solutions Page
http://www.olivesoftware.com/solutions/index.asp
Brief description of the Olive Software function

General Principles for Electronic Scholarly Editions
http://sunsite.berkeley.edu/MLA/principles.html
Peter Shillingsburg's principles from 1993 – pre Web

MLA' guidelines for Electronic Scholarly Editions
http://sunsite.berkeley.edu/MLA/guidelines.html
Preliminary guidelines for editors, publishers, consultants and reviewers

TEI description of text editing
http://www.tei-c.org/Activities/ETE/Preview/mcgann.xml?style=text
Buzzetti and McGann address issues to do with textual editing

Textual Criticism-Scholarly Editing – Wilhelm Ott
http://www.allc.org/reports/map/textual.html
Refers to problems in attitudes to electronic scholarly editing

Summary of DRH 2005 – Alexander Huber
http://www.lib.ox.ac.uk/staff/staffdev/staff_dev/reports/digital_resources_humanities.doc
Quite a lot of material relating to textual editing issues

The Emergence of the Social Text
http://www.erudit.org/revue/ron/2006/v/n41-42/013153ar.html
Jerome McGann on a version of Coleridge's poetry

JODI article on Text Cycles
http://jodi.tamu.edu/Articles/v06/i01/Hillesund/
Electronic publishing, digital publishing and text cycles

Peter Robinson on the State of Electronic Scholarly Editions
http://computerphilologie.uni-muenchen.de/jg03/robinson.html
Where we are with electronic scholarly editions and where we want to be

Modern Languages Association
http://www.mla.org/cse_guidelines
Guidelines for editors of scholarly editions

Notes from Bergen Seminar
http://jilltxt.net/?p=1409
Talks about a range of issues including popularity and economics of critical editions

Peter Robinson on Electronic Publishing
http://www.digitalmedievalist.org/article.cfm?RecID=6
Current issues in making digital editions of medieval texts—or, do electronic scholarly editions have a future?

Beowulf Image Processing Issues
http://www.uky.edu/~kiernan/eBeowulf/ksk-llc.htm
Summary of the early imaging attempts to decipher the Beowulf manuscript

Kentucky, GODDAG extended XPath
http://mustard.tapor.uvic.ca:8080/cocoon/ach_abstracts/proof/paper_174_iacob.pdf
Iacob and Dekhtyar

Systems/Tools

Olive British Library search page
http://www.uk.olivesoftware.com/
search template access to pilot project

Folio Views Search Engine
http://www.nlx.com/pstm/pstmsoft.htm
Software shipped with the Past Masters data

Virtual Lightbox
http://www.mith2.umd.edu/products/lightbox/
Online image comparison tool

Versioning Machine
http://www.v-machine.org/index.php
Software Tool to display and compare texts in multiple versions

AnyText for Mac
http://www.worldlanguage.com/Products/AnyText-for-Mac-English-US-Greek-Classical-Russian-Religious-Studies-Any-1159.htm

Comprehensive Perl Archive Network
http://www.cpan.org/
Perl programming resources

OWL web ontology language
http://www.w3.org/TR/owl-features/
Allows the building of ontologies – frameworks that support applications and define semantic relationships

Fedora
http://www.fedora.info/
Flexible open source tools for managing and delivering content – particularly multiple versions of objects

Anastasia
http://www.sd-editions.com/anastasia/index.html
Analytical System Tools and SGML/XML Integration Applications

EDITION
http://www.sd-editions.com/EDITION/
new more user friendly update to COLLATE and ANASTASIA

Classical Text Editor
http://www.oeaw.ac.at/kvk/cte/index.htm
The word-processor for critical editions, commentaries and electronic publishing
with any number of apparatus - bidirectional text – sigla

XML C Parser and the Toolkit of Gnome
http://www.xmlsoft.org/
Very stable and operable XML parser

Shareware Text Editors

http://www.sharewareconnection.com/titles/text-editing.htm
Range of packages for various text editing tasks

Collex
http://www.patacriticism.org/collex/
A collections and exhibits mechanism for the semantic web

Apache Lucene
http://lucene.apache.org/java/docs/
Open source high-performance, full-featured text search engine library written entirely in Java

IVANHOE
http://www.nines.org/tools/ivanhoe.html
shared, online playspace for readers interested in exploring how acts of interpretation get made and reflecting on what those acts mean or might mean

JUXTA
http://www.patacriticism.org/juxta/
cross-platform tool for collating and analyzing any kind or number of textual objects

TuStep
http://www.zdv.uni-tuebingen.de/tustep/tdv_eng.html
Introduction to TuStep text processing software

Anastasia
http://anastasia.sourceforge.net/
Publishing tool for XML/SGML documents

SAXON
http://saxon.sourceforge.net/
XSLT processor

XML/TEI

Slides about Inference Licensing and Markup
http://www.gca.org/attend/2000_conferences/Extreme_2000/Papers/Sperberg-MccQueen/mimslides.htm#x43
Sperberg-McQueen straw man proposal for prolog system to give meaning to markup

Gentle Introduction to TEI
http://www.tei-c.org/P4X/SG.html
XML version of the TEI guidelines

Using XSL and CSS together
http://www.w3.org/TR/NOTE-XSL-and-CSS
W3C information page

TEI Overlapping Hierarchies SIG
http://www.tei-c.org.uk/wiki/index.php/SIG:Overlap
List of proposals to deal with overlapping hierarchies

OASIS on overlapping hierarchies
http://xml.coverpages.org/hierarchies.html

Summary of articles on overlapping hierarchies in XML