The Digital Humanities and textual scholarship. Integration between digital resources in the humanities, challenges and possibilities

Spence, Paul
King's College. Centre for Computing in the Humanities. 2nd Floor.
26-29 Drury Lane. London. WC2B 5RL
paul.spence@kcl.ac.uk

BIBLID [1137-4454 (2010), 25; 139-148]
Recep.: 24.04.2009
Accept.: 22.09.2010

Artikulu honek produkzio akademiko digitalaren eginkizuna aztertzen du humanitateen alorreko testuen moldaketa, edizio, kontsulta eta argitalpenari dagokionez. Bereziki aipatzen dira goraka ari den "Humanitate Digitalak" alorra eta horien babesean sortu diren esparruak eta arauak, Text Encoding Initiative (Testuak Kodetzeko Ekimena) eta arlo bereziko ekimenak barne, zeinek ikerketaren eta edizio digitalaren arteko integrazio handiagoa bermatzeko helburua duten.

Giltza-Hitzak: Testu digitala. Humanitate Digitalak. Erudizio digitala. Text Encoding Initiative (TEI). *Testu-modelizazioa. Ikerketa-metodologiak. Ikerketa integratua. Humanitateen informatika.*

Este artículo explora el papel de la producción académica digital en el modelado, edición, consulta y publicación de textos de humanidades. Se hace especial referencia al campo emergente de las "Humanidades Digitales" y de los marcos y normas que han surgido al amparo de las mismas, incluida la Text Encoding Initiative *(Iniciativa de Codificación de Textos) y de las iniciativas de dominio específico para garantizar una mayor integración entre investigación y edición digital.*

Palabras Clave: Texto digital. Humanidades Digitales. Becas digitales. Text Encoding Initiative (TEI). *Modelado de textos. Metodologías de investigación. Investigación integrada. Informática para las Humanidades.*

Cet article explore le rôle de la production académique digitale dans la modélisation, l'édition, la consultation et la publication de textes en sciences humaines. Il fait notamment référence au domaine émergent de les "Sciences Humaines Digitalisées" ainsi qu'aux cadres et aux normes qui ont émergé sous son égide, dont la Text Encoding Initiative *(Initiative de Codification de Textes) et des initiatives spécifiques aux domaines pour assurer une plus grande intégration dans la recherche impliquant l'édition digitale.*

Mots-Clés : Texte digitale. Sciences Humaines Digitalisées. Production académique digitale. Text Encoding Initiative (TEI). *Modélisation de textes. Méthodologies de recherche. Recherche intégrée. Informatique des sciences humaines.*

Many claims have been made about the benefits of digital scholarship in the last few years, including the idea that it can widen accessibility, enhance existing scholarly practice with new methodologies and tools, and ensure greater integration between research. The reality has rarely lived up to the 'hype', and for most scholars digital tools and methods provoke as many questions as they provide answers. There are, nonetheless, some clear benefits to this key area within the 'New Humanities', and the issue is not whether or not to use digital methods, but rather when, and how.

The loosely defined field of the 'Digital Humanities' aims to engage precisely with these questions, and has been instrumental in establishing methods and tools to ensure that digital scholarship is underpinned by both technical and non-technical standards. This paper explores the mediating role of the Digital Humanities and its significance for the edition of texts, setting out some of the questions that we need to answer when we decide to use computers for textual scholarship, identifying some of the principles which lay the foundation for emerging 'communities of best (digital) practice' across the humanities and examining possibilities for integrating knowledge from different research areas.

I work at the Centre for Computing in the Humanities (CCH), an academic department within the School of Arts and Humanities at King's College London which fosters the use of computing in research and teaching, and which studies its impact on scholarship across the arts, humanities and social sciences. CCH has a long history of collaboration with other European partners, including Carmen Isasi and other colleagues at the University of Deusto who have carried out pioneering work in the area of text-based digital research, and it was a particular honour to attend the seminar hosted by Eusko Ikaskuntza on new perspectives in the edition of Basque texts.

One thing that struck me while preparing for my talk in Bilbo in December 2008 was how much my own department, and for that matter the field of Digital Humanities in general, has in common with the features of Eusko Ikaskuntza described on its website, namely its internationalism, interdisciplinarity, flexibility and stated desire to bring together institutions and professions[i]. Digital Humanities is by its very nature highly interdisciplinary, both drawing on and engaging with a wide range of humanities disciplines (from history to linguistics, and from musicology to cultural studies and beyond), as well as wider scholarly and professional research methodologies and standards[1]. In the more than 40 research projects that CCH is currently involved in, a range of researchers (including art historians, palaeographers, linguists, medievalists, classicists, literary scholars and historians) engage with a broad spectrum of digital specialities (including text encoding/modelling, relational database modelling, image digitisation, GIS mapping, 3d visualisation), frequently blending more than one to create integrated digital resources[ii].

1. This has been described in great detail elsewhere, including http://www.iath.virginia.edu/hcs/mccarty.html Accessed 18 April 2009.

There is an important distinction to be made here between the 'service' orientation of some digital practice –which treats humanities researchers as 'clients'– and the collaborative approach of the Digital Humanities, which aims to ensure that humanities research questions drive the technical choices, rather than the other way round. It is a relationship which is collaborative in every sense of the word: the Digital Humanities specialists play a mediating role between humanities domain specialists and the technical research, and it is common for both groups to be transformed by the experience. It is also important to note the necessary fusion of both technical and humanities methods: it is relatively easy to build a database or text corpus, but it is much harder to ensure that they match the academic and technical standards required to make them future resistant and 'interoperable' (i.e. sense of enabling different systems to work together) in the true sense of the word[2].

My own research interests lie in textual modelling and the challenges in representing, editing, querying and publishing texts in digital form and the 'Digital Text' team I manage at CCH has involvement in a wide range of projects, which include research into nineteenth century musical catalogues, Anglo-Saxon charters, classical epigraphy, the reception of Spanish language theatre in the English-speaking world and medieval historical documents.

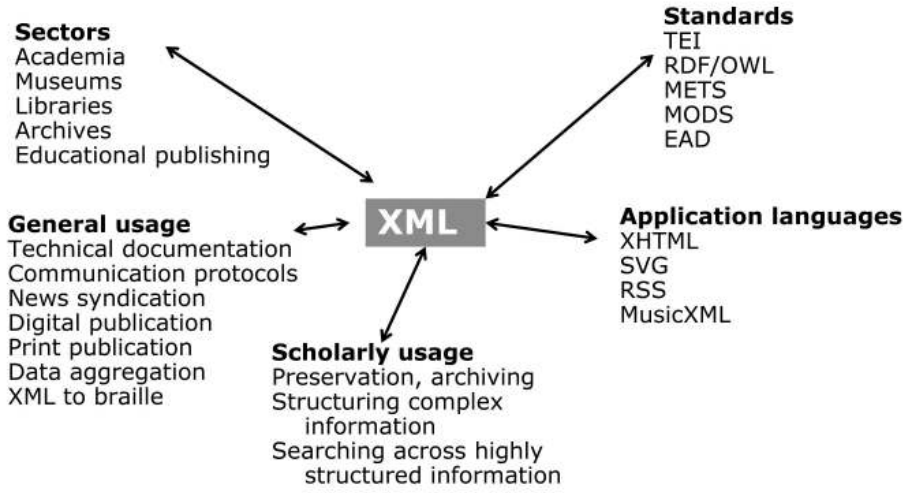
The technology at the heart of much of CCH's work is XML (eXtensible Markup Language)[iii], which has some history in document publishing, but which is extensively used as a platform and software-neutral standard for data interchange, although its presence often goes unnoticed[3]. XML has a wide range of applications in the commercial world including technical documentation, communication protocols, news syndication, digital and print publication and data aggregation. To these more general applications we might add a number of potential benefits specific to academic scholarship, such as its key role as an underlying format which facilitates data preservation and integration between the scholarly community, libraries and museums; or the ability to model complex structural and semantic information about texts. This last point constitutes one of its most common uses in humanities research: briefly put, XML provides scholars with the potential to represent texts in a wide variety of sophisticated ways which separate content and structure from its eventual presentation. We will now explore what this means in practice for textual scholarship in more detail, making loose reference to Allen Renear's analysis of the advantages of descriptive markup in general (and therefore XML in particular)[iv].

---

2. 'Communities of best practice' such as the *Digital Classicist* and *Digital Medievalist* are examples of the kinds of initiatives that strive to make this interoperability reality.

3. In a January 2008 blog post, Liam Quin responded to questions about the impact of XML: 'I was recently asked, how widespread is the use of XML? I had to stop and think. It's almost as if someone asked me how widely used is air, or perhaps more fairly how many socks there are in the world.' http://people.w3.org/~liam/blog/?p=7 accessed 18 April 2009.

XML everywhere (museums, libraries and archives)?



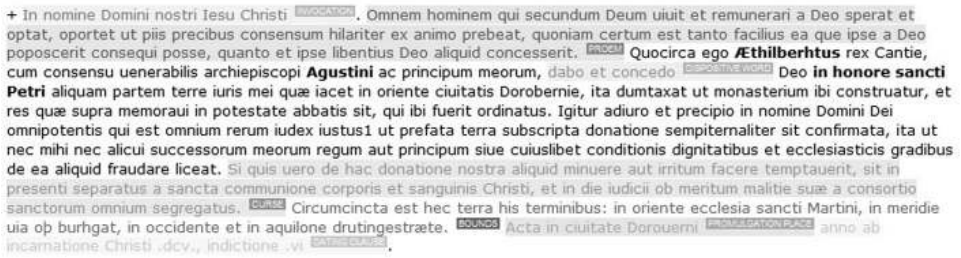
As described by Renear, XML benefits the authoring process in a number of different ways by simplifying the composition process and allowing authors to focus on the structure and semantics of a text rather than its eventual rendition(s). This ensures a consistent data structure that can be 'validated' against any rules we wish to define for a given set of documents, ensures that editing is based on meaningful concepts (headings, text divisions etc) rather than presentational cues provided by an individual piece of software, and makes it possible to devise structured authoring environments specific to a given field, and to provide authors with a number of document views for checking purposes.

On CCH's projects we have developed XML models to allow scholars: to represent diplomatic discourse in Anglo-Saxon charters so that users can compare different formulae used; to model key features in a nineteenth century musical catalogue so that users can browse or query them, and view indices of composer, publishers and publishing places; to make statements about text that is missing or unclear in late antique stone inscriptions. In summary, you can mould an XML-based framework to suit the nature of the material and the scholarly research needs.

XML is particularly useful where you wish to edit a text without making a specific commitment to a single rendition, a feature which is helpful where different scholarly traditions or interpretations make it difficult or risky to do so. In his work to produce a scholarly edition in English of the Cervantes play *La entretenida*, John O'Neill from the Out of the Wings project has, in collaboration with CCH, developed an XML-based encoding scheme for the play

which allows it to be viewed according to a number of different editorial criteria, including scholarly, translational or performative criteria (chiefly aimed at potential directors/actors of the play)[v]. It is an approach which recognises the inherent instability in many (if not all) humanities texts, in this case due to the process involved in producing the play in all its stages, which includes the 'interfererence' of the *amanuensis*, *corregidor*, typesetters) and the innately unstable nature of the performance process, which involves actors, directors and translators. As O'Neill himself states: 'this quality of instability requires a different approach to editing, which does not seek to fix the text, but instead is able to present it in many different ways'[vi].

I have already hinted at some of the advantages for publication: an author can maintain a single set of XML documents, and from this any number of different renditions (or visualisations) may be generated. In one medieval historical project, researchers interested in the English king Henry III edited summary translations of the historical source materials in XML, which were then used to produce digital textual representations of the sources, rich indices containing information about various entities (persons, places and subjects) mentioned in the documents, as well as a structured search facility to query the same entities[vii]. The principle of *single source publishing* loosely followed here minimises the risks involved when the same information is maintained in different places, ensures consistent formatting and facilitates the automation of the kind of auxiliary navigation (tables of contents, lists of figures) that any scholarly edition typically requires. Finally, XML facilitates publication in a number of different formats appropriate to print or digital publication: the *Henry III Fine Rolls* project has produced both print volumes and web resource using the same XML source materials.




```
3  <text>
4
5      <p>+ <invocation>In nomine Domini nostri Iesu Christi</invocation>. <proem>Omnem hominem qui
6         secundum Deum uiuit et remunerari a Deo sperat et optat, oportet ut piis precibus
7         consensum hilariter ex animo prebeat, quoniam certum est tanto facilius ea que ipse a
8         Deo poposcerit consequi posse, quanto et ipse libentius Deo aliquid concesserit.</proem>
9         Quocirca ego <name><donator>Æthilberhtus</donator></name> rex Cantie, cum consensu uenerabilis
10        archiepiscopi <name><beneficiary>Agustini </beneficiary></name> ac
11         principum meorum, [...] </p>
12 </text>
```

Perhaps the greatest scholarly benefits of using XML are in the area of information retrieval and analysis: the relative difficulty in producing (and then using) such query/visualisation tools is probably also one of the reasons why this is one of the least realised benefits of the technology in practice, but it is possible to create highly sophisticated and context-aware structured query tools or to produce complex statistical data based on the structure of a document.

Finally, two related benefits that we might add to those described by Renear include the fact that XML facilitates preservation and 'interoperability' (hence its status as a standard for data interchange) and integration: a good example of this is the collaborative project *Integrating Digital Papyrology*, which aimed to create a multi-institutional, international database of papyrus collections, and where XML was used at various stages not only as an authoring and preservation format, but also as an underlying format to integrate scholarly data from some of the foremost institutions in papyrology. I will return to the issue of integration in greater detail later.

As we have seen, XML allows us to represent a wide variety of materials, to visualise them in different ways and to do so in a manner which ensures some level of interoperability, but this interoperability becomes much more real when researchers follow similar models/guidelines as they deal with similar challenges. The *Text Encoding Initiative* (TEI) was developed precisely for this reason:

> The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics[viii].

TEI has become an XML-based *de facto* standard for the digitisation of text in the humanities and represents an example of one of the oldest and largest collaborative humanities research projects ever carried out. The TEI community brings together people from libraries, museums, publishers and scholarship and provides a framework for people both to use and contribute to when carrying out text-based digital research. The TEI guidelines provide recommendations and encoding samples for most scenarios a humanities researcher is likely to face, the user is provided with a rich set of modules that can be applied according to their own particular requirements and the latest version of TEI, 'P5', has made the process of creating your own encoding scheme, re-editing it and then documenting it much easier.

There is an active TEI mailing list which allows people to ask for and provide advice[ix], as well as more specialised working groups and Special Interest Groups, which focus on individual challenges or fields associated with text encoding (how to encode information about people; correspondence; manuscripts; text and graphics; education). There are also a number of loosely connected TEI-based initiatives specifically designed for particular scholarly needs: a good example is the EPIDOC initiative for encoding epigraphic materials[x].

TEI is often used to encode digital editions of primary texts, or associated scholarly materials, but humanities researchers frequently require other structures to represent other aspects of their research activity, in particular two areas that I will mention briefly now. The first issue has to do with how we represent different levels of scholarly interpretation when applied to a digital text –TEI is commonly used to represent a digital interpretation of the content and structure of a document, but it is not always so easy to use TEI to represent scholarly statements at a higher interpretative level: for example, when we wish to state that a person mentioned in one document is in fact the same as a person mentioned in another. In the *Henry III Fine Rolls* project, this interpretative layer has been crucial in allowing us to develop the sophisticated indices and search function, which are a key research outcome. There are a number of technical options if this needs arises: for this project we chose a fairly demanding option based on the RDF/OWL technologies associated with the concept of the semantic web[xi], but we could easily have chosen a number of different options (most of them XML-based) – the key criterion was that whatever approach we took, it had to be standards-based.

The other thing to mention here is the issue of metadata. Metadata, literally 'data about data', is too broad a topic to cover in any detail here, but it can be broken down into various categories, including: descriptive metadata (to facilitate access to material)[4], preservation metadata (to facilitate preservation and management of information within a digital repository), administrative metadata (to document the life cycle of a digital resource) and structural metadata (to describe the internal structure of a set of materials). One of the most powerful frameworks for bringing these different kinds of metadata together is the XML-based METS (Metadata Encoding and Transmission Standard) framework: 'The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library'[xii]. While, as the name suggests, METS was developed within the library world, it can be useful for any cataloguing activity involving source materials and scholarly analysis.

Most of the standards and initiatives I have described aim to facilitate interoperability in some way, and most of them use XML as their underlying data format. Since XML is a standard for data interchange, you might expect XML-based research to lead to greater integration of research data in the academic world (data integration is certainly quite common in the commercial world), but in practice there is little evidence to support this idea: the potential is certainly there, but this potential is rarely realised. Part of the reason for this may be scholarly conservatism: exemplified by the response a colleague of mine received when she explained the potential of using XML to broaden access to and integration between epigraphic research: apparently the respondent in question expressed concern that people actually 'might read

4. Examples include MARC, Dublin Core and MODS, largely library-focused initiatives.

her research' as a result[5]. Certainly some digital projects have been guilty of replicating this conservatism in many cases: witness the 'data islands' or 'digital silos' which have been developed over the years, often representing technical research carried out in isolation and without reference to similar research, with the result that much of it is hidden from view, inaccessible to most people except under very specific circumstances and therefore difficult to connect within wider digital research frameworks. But there are also some very real issues, both technical and intellectual, which make integration a real challenge. I will now describe a case study for integration between Anglo-Saxon projects which explores some of these issues.

CCH has been involved in a number of digitisation projects involving Anglo-Saxon studies in the last few years. They include: the *Proposography of Anglo-Saxon England* project (PASE)[xiii], a database providing structured information on all of the recorded inhabitants of England from 597 to circa 1100; *Anglo-Saxon Charters* (ASChart)[xiv], a pilot project to explore the potential usage of XML to represent diplomatic discourse in Anglo-Saxon charters*; Electronic Sawyer*[xv]*,* a revised edition of Peter Sawyer's seminal 1968 catalogue of extant Anglo-Saxon charters; and the *Language of Landscape* (Langscape)[xvi], an on-line searchable database of Anglo-Saxon estate boundaries, descriptions of the countryside made by the Anglo-Saxons themselves.

There is clearly considerable scope for contrasting and connecting the research data from the four projects and since they are all projects in which CCH has been involved, and employ similar technical standards –and, to a certain extent, some basic common technical approaches– the task is somewhat easier as a result. In fact, some superficial integration between the projects has been implemented: for example the links from a representation of a charter in the Anglo-Saxon pilot project to information about the witnesses and sources for that charter in the PASE proposography[6].

There have been a number of proposals to pursue the idea of integration further, and a number of papers exploring the central issues, but it was in a paper session given at the Digital Humanities 2006 conference[xvii] that Paul Vetch, John Bradley and I outlined the key challenges, both on the technical and the non-technical side. One of the central issues we identified was how to provide clear technical systems of identification between resources which take slightly different technical approaches –some of the projects use relational database technology, while others use document-focused TEI XML as the representational layer, and then how to signpost to the user the fact that

5. More serious treatment of this question is provided in *Our Cultural Commonwealth*. The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. American Council of Learned Societies. 2006, p28. Electronic version at http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf accessed 18 April 2009.

6. See, for example: http://www.aschart.kcl.ac.uk/content/charters/text/s0002.html accessed 18 April 2009.

data belongs to one project or another. There are both technical challenges (such as how to provide canonical references to visualisations of underlying data that may be dynamic, and which are not 'stable' documents as such, but rather the result of a user-driven query) and humanities challenges (how to reconcile the fact that one project may not wish to simply 'accept' another project's interpretation of a given set of facts without any kind of annotation or filter).

We explored a number of different approaches, ranging from a loose 'web portal' approach (which might allow information to be brought together without having to deal with issues of scholarly interpretation and authority), to a straight integration of data between different projects (which could allow for more meaningful connections to be made, but carries with it all the challenges described above). It was partly this research that led to the *Anglo-Saxon cluster* proposal, which is currently in progress as a one-year JISC-funded research project to explore possibilities for integration between Anglo-Saxon projects, including (but not limited to) the CCH projects mentioned already.

The *Anglo-Saxon cluster* project brings together researchers from King's College London, Cambridge University and Oxford University and aims to provide a single entry point to various digital resources focusing on the study of medieval Anglo-Saxon materials, with thematic connections between projects, a platform for global searches on data contained in all featured projects, indices which bring together data common to all the projects (which we are currently calling 'union indexes') and options to visualise data generated within a particular project context. We are aided to some extent by the fact that there is a commonly accepted system for identifying charters from this domain: the so-called *Sawyer number*, which Anglo-Saxon scholar Peter Sawyer created in 1968 to catalogue each of the charters which are known to have existed, but our objectives are to explore common methods for actually encoding charters themselves (building on research in the *ASChart* project), to define methods for aggregating research from each of the projects, to define methods for allowing a project to expose its research data for use by other projects not initially covered by the scope of the *Anglo-Saxon cluster* and to publish this information in a public web resource.

While it is not realistic to merge different scholarly interpretations seamlessly, it is undoubtedly of great use to the wider community if they are brought closer together. Connecting research in this way allows scholars to move quickly from one research environment to another, makes the scholarly process more fluid and, of course, facilitates the greater dissemination and juxtaposition of different research outcomes. The technical models created to mediate between source materials and their scholarly reception are crucial to this enterprise, but as we have seen, the use of standards and the emergence of communities of best practice in a given domain are invaluable factors in the development of these scholarly resource networks.

BIBLIOGRAPHIC REFERENCES

i. http://www.eusko-ikaskuntza.org/es/quienessomos/ accessed 18 April 2009.

ii. http://www.kcl.ac.uk/cch/research/ accessed 18 April 2009.

iii. http://www.w3.org/XML/ accessed 18 April 2009.

iv. RENEAR, Allen H. "Text Encoding" In: A Companion to Digital Humanities, 2004. Oxford: Blackwell, 2004; pp 222-224. Electronic version http://www.digitalhumanities.org/companion/, accessed 18 April 2009.

v. http://www.outofthewings.org, accessed 18 April 2009.

vi. O'NEILL, John. 'Editing Cervantes's Plays in the 21st Century: An Experimental Approach to an Experimental Dramatist' – presentation from *Papers on the literature, history and visual arts of the Spanish Golden Age (1474-1681)*, 21 November 2008 at Department of Hispanic Studies, University College Cork.

vii. http://frh3.org.uk accessed 18 April 2009.

viii. http://www.tei-c.org/ accessed 18 April 2009.

ix. http://listserv.brown.edu/archives/cgi-bin/wa?SUBED1=tei-l&A=1/ accessed 18 April 2009.

x. http://epidoc.sourceforge.net/ accessed 18 April 2009.

xi. http://www.w3.org/2001/sw/ accessed 18 April 2009.

xii. http://www.loc.gov/standards/mets/ accessed 18 April 2009.

xiii. http://www.pase.ac.uk/ accessed 18 April 2009.

xiv. http://aschart.kcl.ac.uk/ accessed 18 April 2009.

xv. http://www.esawyer.org.uk/ accessed 18 April 2009.

xvi. http://langscape.org.uk/index.html accessed 18 April 2009.

xvii. BRADLEY, John; SPENCE, Paul; VETCH, Paul, 'Joining up the dots: issues in interconnecting independent digital scholarly projects' session at *Digital Humanities* conference, 2006. Paris-Sorbonne, 2006.