Documentation and the users of Digital Resources in the Humanities

Claire Warwick, Isabel Galina, Jon Rimmer, Melissa Terras, School of Library, Archive & Information Studies, University College London, Gower Street, London WC1E 6BT	Ann Blandford & Jeremy Gow, Interaction Centre, University College London, 31-32 Alfred Place, London WC1E 7DP	George Buchanan Future Interaction Technology Laboratory, Computer Science, University of Wales, Swansea SA2 8PP

Corresponding author

Claire Warwick

School of Library, Archive & Information Studies,

University College London,

Gower Street, London WC1E 6BT

+44 20 7679 2548
c.warwick@ucl.ac.uk

Documentation and the users of Digital Resources in the Humanities

Abstract

Purpose

This article discusses the importance of documentation for digital humanities resources. This includes technical documentation of textual markup or database construction, and procedural documentation about resource construction.

Methodology

We present a case study of an attempt to re-use electronic text to create a digital library for humanities users, as part of the UCIS project. We discuss the results of qualitative research by the LAIRAH study on provision of procedural documentation, and user perception of the purpose, construction and usability of resources collected using semi-structured interviews and user workshops.

Findings

In the absence of technical documentation, it was impossible to reuse text files with inconsistent markup (COCOA and XML) in a Digital Library. Also, although users require procedural documentation, about the status and completeness of sources, and selection methods, this is often difficult to locate.

Practical implications

Creators of digital humanities resources should provide both technical and procedural documentation and make it easy to find, ideally from the project website. To ensure that documentation is provided, research councils could make documentation a project deliverable. This will be even more vital once the AHDS is no longer funded to help ensure good practice in digital resource creation

Originality/value

Previous work has argued that documentation is important. However, this paper presents actual evidence of the problems caused by a lack of documentation and shows that this makes reuse of digital resources almost impossible. This is intended to persuade project creators who wish resources to be reused to provide documentation about its contents and technical specifications.

Keywords

Documentation, digital libraries, digital humanities, XML markup, humanities users, re-use of data.

Classification: Research paper

Introduction

It has long been agreed in computing that when a resource, a program or code is being created, it ought to be documented. (Raskin, 2005) As AHDS (Arts and Humanities Data Service) History explain: “Good documentation is crucial to a data collection's long-term vitality: without it, the resource will not be suitable for future use and its provenance will be lost. Proper documentation contributes substantially to a data collection's scholarly value.” (AHDS History, no date)[1]

Many research projects in the arts and humanities now have some form of digital output. Most of this digital content is in the form of electronic text, or websites which enable a researcher to query and navigate through a dataset or image database. Much of the textual data has been deposited in text archives, and is marked up, for example in SGML or XML (Standard Generalised Markup Language and eXtensible Markup Language, respectively). Such markup is non-platform-specific, and thus not tied to any software or hardware, and is designed to make reuse of resources easier, especially when the user may have no knowledge of or access to the original resource creator (Morrison et al, no date, chapter 4).

The ideal textual digital resource deposited in an archive should therefore be an XML[2] document, carefully and consistently marked up, accompanied by detailed documentation of both the markup principles and the history and provenance of the digital text. Likewise, a web based digital resource which contains research material, such as a database, or image repository, should have careful and consistent documentation regarding the purpose, provenance, and coverage of the site's contents, and the technical details of implementation necessary for users (and future re-use of content). However, this is not always the case. Resource creators may begin with good intentions, but documentation may be left late, done in a hurry, or not done at all. (Warwick, et. al. 2007)

The following article therefore describes the problems that may emerge when documentation of digital resources is neglected. It reports on the research of two related projects, UCIS and LAIRAH, whose research studies the way that humanities scholars interact with digital libraries and digital research resources. Through a case study of our experience on the UCIS, (User-centred Interactive Search with Digital Libraries) project [3], we demonstrate why documentation, commenting code and the accurate use of SGML and XML markup are vital if digital resources are to be reused. The LAIRAH (Log Analysis of Information Resources in the Arts and Humanities)[4] project carried out research on the reactions of users to digital resources and their usability. We investigated the problems users have encountered in finding information about digital resources in the humanities, and how this has affected their perception of the resource's quality and possible usefulness in their research.

Our findings are even more relevant since the decision (announced in June 2007) that the AHDS would no longer be funded. The problems that we report affected files which had been created before the existence of a central archiving body which also provided advice on good practice in data creation; they also show a wide variety of practice in the use of markup conventions. If such a central body no longer exists there is every chance that differences in the way that digital resources are created and documented will once again proliferate, making reuse ever more difficult. The research presented below therefore demonstrates that documenting what producers do when creating digital resources is now even more important.

Background to the Projects

This article represents a synthesis of the research of the UCIS and LAIRAH projects. The UCIS project is studying the way that humanities researchers interact with digital library environments. We aim to find out how the contents and interface of such collections affect the way that humanities scholars use them, and what factors inhibit their use. (Warwick, et al., 2005) An early work-package of the project was to build a digital text collection for humanities users, delivered via the Greenstone digital library system. However this task was to prove unexpectedly difficult. Below we describe the problems that we faced in our attempt to reuse texts which we had intended to include in the Digital Library system.

The LAIRAH project studied the factors that influence the long-term sustainability and use of digital resources in the humanities through the analysis and evaluation of real-time use. We attempted to discover whether digital resources which appeared to be neglected might be re-used, by introducing a sample of used and neglected resources to potential users. We also studied a sample of well used projects with the aim of discovering if they shared any elements of good practice. One of the issues that we researched was that of the documentation of digital resources, and how this affects their usability.

Documentation Defined

Our study relates to use of digital resources in the humanities, but the issue of how to document decision making is of interest in a wider range of fields. Documentation might be described as extended metadata. Hugh Denard and his team have called it “paradata”, in other words text that provides information about decisions taken in the construction of a digital resource, and the reasons for them. (Beacham et al, 2006)

Technical and Procedural Documentation

Documentation has two different functions, and thus can be found in different forms, which we have termed technical documentation and procedural documentation. In the

following article we describe in detail the use of both types. Technical documentation is the kind of detailed information that would accompany a collection of marked up documents, or a database. It should describe the rationale for the creation of the markup scheme and the way that individual tags are applied, or function as a detailed codebook for a database, showing how fields were allotted and explaining relationships between fields. This allows anyone who wishes to reuse the data to understand the rationale for its creation, and is especially useful if there are any problems with the consistency of how the scheme has been applied, since it allows problematic data to be reconstructed.

Procedural documentation is more general and describes the way that the project itself was created and run. It should include details of the data sources, especially if they were digitised from analogue material. It should show whether the collection was comprehensive, and if not, how material was selected and why. It should detail any important decisions taken in the running of the project and creation of the resource. In effect then it preserves the institutional memory of a project. Technology changes with time, and thus solutions developed by an earlier project may become outmoded. However, if new projects can consult the documentation produced by others, they may be able to adapt existing resources or discover solutions to similar problems, and thus could save significant amounts of time and money. This is what Orna (2005) describes as the process of making tacit knowledge visible, and she argues that it is vital to any successful organisation, since without this process information is lost, and has, in effect, to be continually recreated. Procedural documentation also enables users to access as much information as possible about the contents of the resource, and the decisions taken in its construction. As we will see, this helps give users confidence in the quality and reliability of digital resources.

Structured Documentation

The documentation produced by one research team may make little sense to another scholar unless it is structured in some way. Research on health information, for example, has shown that the attempt to impose structure on human-generated information is fraught with difficulty. The attempt to produce systems for electronic patient records such as the NHS's Information for Health initiative, (NHS 1998) is a

huge documentation problem. If doctors wish to be able to share information about what the patient has told them, and the decisions and diagnoses that they have made, then there must be some agreement on the structure that these records ought to take. However, numerous studies have shown that even within set fields, different health professionals enter different information. This hinders interoperability, and necessitates complex training and monitoring of staff. (Porcheret et al, 2004). Walsh (2004) argues that problems come about because the nature of the interaction between doctor and patient is that of capturing a patient's narrative, and that such a narrative does not fit well into the kind of structured data fields to be found in medical records systems. Yet the more detailed free-text that can be entered, the more difficult it may be to cross-search large systems.

This problem has also been encountered in the attempt to assign metadata to digital resources. Dublin Core metadata, for example, is simple to apply because is has only 15 fields into which the user can enter free text data. (Dempsey and Weibel, 1996) It was designed to allow non-experts to document the web publications that they created. (Miller, 1995) Different interpretations of the contents of each field and the level of detail required may limit cross searching, and this quickly led to the development of qualifiers for each field, making the scheme more precise, but at the same time more complex. (Knight, 1997) The alternative is the use of controlled vocabulary. However, this usually requires the kind of training only given to library professionals, and would immediately deter most data creators who are not information specialists. (Heery, 1996) The problematic choice therefore in documenting digital resources is between free text metadata, documentation that may be variable in extent and quality, or perhaps the lack of any documentation at all.

Folksonomies have been suggested as a solution to the problems of inconsistency in provision of metadata. Users of a resource, rather than its creators, could create a tag cloud using an application such as del.icio.us, or Connotea to describe a resource . (Voss, 2007) However, in some ways this multiplies the problems of metadata which lacks controlled vocabularies, since taggers may apply a large number of different terms for the same feature. (Kipp & Campbell, 2006) It may therefore lead to greater confusion amongst users. As we shall show below, users of digital humanities resources value authoritative information about resources, which reassures them of its

serious scholarly content. They also liked to know that the creators of resources were well qualified, and that content could therefore be trusted. The more informal, community based approach of folksonomies may therefore lack the requisite authority and scholarly rigour to be trusted by users of academic resources for the humanities, at least at this early stage of their development.

The problem of which terms to choose for metadata is immediately relevant to electronic text and markup of the kind that we describe below. XML markup in particular is flexible by design. Although it must be well-formed, which means that tags should be used consistently, an XML document can be used without even a schema or DTD (Document Type Definition). Furthermore, SGML and XML allow the user to mark up the text using any tag that they wish. (Sperberg McQueen and Burnard, 2002) Thus <p>, <pg> <page> may all be used to indicate a page in different markup schemes.

Many texts used in digital humanities are marked up in TEI (Text Encoding Initiative) XML, which was developed to produce a more standardised tag set for humanities texts.[5] (TEI, 2001) (Hockey, 2004) Yet TEI is now a huge scheme, and it is also relatively common for users to add their own customizations to the scheme. (TEI Consortium, 2007, Chapter 29.) It has long been realized that adding markup is an act of interpreting a text, and that the way that different encoders mark up will depend on their interest in the material itself. (Sperberg McQueen, 1991) If using TEI, the basic structure of the markup will probably be the same, with divisions, paragraphs, line numbers etc. Yet any more complex content markup will tend to differ depending on the judgment of who is producing the text. Even when using the same DTD on the same material, a team of encoders may differ in their markup. (Butler et al., 2000) W3C schemas allow for very detailed data typing, which may help to eliminate the problem, but may also be beyond the capability of many encoders to design. Similar problems exist when different coders produce metadata for the TEI header about the document history and development. (Daneker and Warwick, 2005) Theoretically, then, markup is designed to aid interoperability, yet the freedom allowed to those designing, and indeed applying, markup schemes may militate against this.

UCIS and the Re-use of Digital Text

In this section we describe our initial examination of the files in the UCIS case study. We then go on to discuss more general issues relating to documentation that were brought into particular focus by the experience.

As part of the UCIS research we aimed to build an interface to the Greenstone digital library system (Witten et al., 2001) designed to meet the needs of humanities users. As a result a text collection was needed, to populate the digital library. We chose to use texts from the Oxford Text Archive, (OTA) [6] because this substantial collection is freely available and contains at least basic levels of XML markup. We therefore contacted the OTA, who sent us a varied sample of texts, many of which were relatively old, and thus not produced under AHDS guidelines. (The OTA was part of AHDS Literature Language and Linguistics). Initial problems with the sample texts came about partly because we were not able to find any documentation to accompany the text files. It was therefore necessary to examine them in detail to determine whether we could reconstruct the rationale for their markup. On examination of a sample of the files, we found that although they appeared to be in well formed XML, there were many inconsistencies in the markup.

However, although the problems that we report below may appear extreme, the texts are still available, and thus other potential users are liable to suffer from the same problems. There is also no indication to users when texts are downloaded about the state of the markup that they are likely to encounter. The reason for deposit with the AHDS was to facilitate the long term re-use of digital materials. It is therefore possible that what seems best practice in markup now may seem equally outlandish to users in twenty years time. Thus it is instructive to examine the problems in the most difficult texts, since these are not museum pieces, but items intended for some future use.

These texts also demonstrate the problems caused when different types of local practice are used in text markup. This situation was undoubtedly improved by the advice given by the AHDS, especially to projects funded by the AHRC. However, in

future, resource creators will not be able to access such advice, and given varying levels of access to support for digital humanities, it seems likely that such variations in practice will again proliferate, except in those few universities that are served by specialist humanities computing centres.

Markup Problems

Inconsistencies often arise from the electronic history of documents. In our sample, problems were caused both by clashes between different markup schemes, and the way that these schemes were applied by text creators. Many of the texts in the sample were older (Early and Middle) English texts, whose structure is complex, and many of the problems stemmed from succeeding revisions to the underlying content. One common early standard was COCOA markup, and many of the documents still contain COCOA tags which meant that the files would not parse as XML.

COCOA is a markup scheme that preceded XML and SGML. Unfortunately, it can look, both to humans and machines, just like XML, as markup is contained in angle brackets (< >). However there are important differences. In XML tags normally appear in pairs, which encompass a portion of document content, whereas in COCOA tags are typically singletons that mark a particular point in the document. The internal structure of COCOA and XML tags is also different. COCOA tags consist of two parts, one indicating the name of the tag and the other its value, for example <line one>. In XML this information would be provided as an attribute value <line id= “1”> (Frackoviak, 1999). In our sample texts, the tags were retained in their original COCOA form which was mistaken for potential TEI tags by the processing software.

In addition, COCOA uses different notations for extended characters than XML, and there is variance even between COCOA documents. Many letterforms found in earlier English were encoded in OTA documents using idiosyncratic forms where modern (Unicode or SGML Entity) alternatives now exist. For example, one common character notation was ‘&&’ to represent the ‘Thorn’ character (in upper case) and ‘&’ to represent the same character in lower case. This was interpreted as an SGML/XML entity, but parsers were unable to successfully interpret the original scheme. Furthermore, as the SGML/XML format was used in other parts of the

document, even a bespoke parser could not successfully disambiguate the intention of every occurrence of the '&' character. Other characters then had to be rendered in forms such as '%' for '&' or 'and' – because of the original special use of '&' to indicate a Thorn. This further complicated the processing of the COCOA/TEI markup. Such characters thus remain unintelligible to an SGML or XML document reader. The earlier, COCOA, form may render the modern electronic encoding unparsable in either XML or SGML.

Furthermore, human encoding proved inconsistent. This is a particular problem since COCOA markup was never fully standardised, and tags are often created or used idiosyncratically. (Lancashire, 1996) This complicates a number of potential technical solutions (e.g. the use of XML namespaces). Some content included unique tags such as "<Cynniges>":(To denote the Anglo Saxon word for 'King') not part of any acknowledged hybrid of the original standard. The purpose of this is unclear. It may be an original part of the text, (words actually surrounded by '<' and '>'), a COCOA tag, or a TEI/SGML/XML tag. In other cases, '<' and '>' are clearly used as syntactic markers within the text, as well as elements of COCOA tags. Thus the precise distinction between COCOA and SGML/XML was not possible in this context. Even parts of the same document used the same tag inconsistently. For example, distances (e.g. "ten lines of space") might be rendered in numeric form ('10 lines') or textual form ('ten lines') and distance units might be given in full or abbreviated form.

The use of white space and implicit formatting was a common feature of OTA content. Heading text was often not distinguished in any way from body text, and traditional plain text formatting was retained, with the explicit marks of line ending one would expect from an XML/SGML document being absent.

Where singleton tags were used in TEI/XML, these did not use the trailing '/' now required for tags that have no corresponding closing tag. (e.g. <pb id="25" /> to indicate page 25) Despite the help of the OTA the semantics that would be explicit in properly formed XML were made potentially ambiguous by incomplete application of modern markup. Page tags provide multiple examples of these.

In XML, values should be encoded as attribute label/value pairs within tags (e.g. value= “example”). In COCOA, tags were written with the value alone – the attribute label being assumed from the tag name. OTA documents often followed the COCOA practice of presenting page breaks as a point in the text. Thus page tags appeared in the form ‘<p 25>’. In contrast, XML/TEI would require the value ‘25’ to appear in a value/attribute pair: ‘<pb id=“25”>’. While in XML the attribute label may be used on its own, implying some default value, any value *must* be accompanied by a label.

Even this modified representation is in error. When parsed as XML, the lack of a closing tag (‘ ’) to partner an opening tag (‘ ’) would produce an error. If one maintains the original semantics in which page tags were encoded as singletons, the proper XML syntax would be ‘<pb id=“25” />’. Note the closing ‘/’ used in singleton tags in XML. However, this approach would mean that the page text was not treated as child content of the page tag. Perhaps a more complete XML representation would be to enclose the page content in an open/close tag pair: ‘<pb id=“25”>…page 25 content…</pb>’, however this would not be usual under TEI guidlelines. Clearly, this representation is now very different from the <p 25> singleton found in the original, nominally TEI, document.

In the case of page numbers, we can intuitively resolve the issue of whether a tag is semantically a singleton or spanning tag. However, in the case of unfamiliar or unknown tags, the appropriate treatment is not immediately clear – exact knowledge of the tag semantics is required. The variation of practice within even a single document, and the apparent confusion between tag markup and document content (see Cynniges above), made it almost impossible to automate the data processing, since it was often impossible for a computer to determine the proper form of the document without human intervention, making the total effort required across a large number of documents very high. Had the markup conventions and semantics of tagging been made specific in accompanying documentation the situation might have been very different. Even if application of the markup had been inconsistent, we would have known what the creators meant to do, and could therefore have corrected the inconsistencies ourselves.

Instead, despite the generous help of Oxford University staff in cleaning up the data, the task ultimately proved so large that we had to abandon the use of these files. We have therefore used commercially produced resources to complete the research, with the permission of their publishers. The advantage of using this material was that the markup is more consistent, has been documented and conforms to written specifications. Thus although we have found that the markup of some files has been inconsistent, we knew what rules it should have conformed to, and were able to adapt the organisation of the data accordingly.

Documentation and the TEI Header

As Giordano (1995) argues, 'No text encoded for electronic interpretation is identifiable or usable unless it is accompanied by documentation'. Yet as we have shown, none of the markup decisions had been documented, nor was the code commented. The OTA supplied each file with a TEI header, which provides some basic metadata about its creation. However, the header was intended to act as the kind of metadata that aids in resource discovery, rather as code books were used to find a dataset on magnetic tape. The <encodingdesc> element is not mandatory, and was intended to explicate transcription practices rather than detailed markup decisions. (Giordano, 1995) We did not find any examples of attempts to elucidate markup schemes in the headers.

Though the OTA strongly encourage depositors to document their work, they do not mention markup specifications as an element of basic documentation, so even documented files might not have provided the information we needed. (Popham, 1998). This situation has changed somewhat with the release of TEI P5 (TEI Consortium, 2007). The TEI header now allows for more detail about the encoding scheme to be included and the ROMA software now helps users produce a schema and an ODD file-(One Document Does it all) containing detailed documentation of the markup. (Sperberg-McQueen and Burnard, 2006) Nevertheless, this ODD file is designed for XML experts, and while generated in HTML and literally human readable, would not be comprehensible to anyone who did not have a reasonably good understanding of XML markup and TEI in particular. Although it may serve as de facto documentation for other technical teams, it was not designed to serve this

purpose. It would not be of use to most future users, or non-technical would be project creators. The most serious problem, however, is that it is possible to use the TEI scheme without generating or including an ODD file, and thus some projects may decide not to use it at all.

Documentation and the AHDS

The AHDS encouraged documentation, and produced a guide to what should be included. They recommend that the structure of the resource should be described, including:

> - List of files and tables with information about their contents, number of records and fields, and the way in which they relate both to each other and to the source
>
> - List of field names used in each file with information about the characteristics of each field, including name, contents, field length, data type and any codes used, and information about the way in which the fields relate to each other and to the source, including details of derived variables (James, 2004)

Perhaps because many of their holdings are the kind of large data sets used in archaeology and History, the AHDS guide appears predicated on a database model of delivery. The instructions above could be applied to markup, but there is no specific mention of markup conventions, and thus some depositors may not have understood that this was needed. Although the AHDS attempted to work with users to improve the quality and complexity of documentation, this could be a demanding and difficult process. (Dunning, 2007)

The AHRC has now recommended that resources should be archived in institutional repositories. Such repositories use the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) which mandates Dublin Core metadata as minimum requirement, but the systems can add other metadata as well, depending on the requirements of each institution. Some institutional repositories let the depositors fill in the Dublin Core fields with no further checking, others have someone who checks it and others do all the metadata centrally. (Probets and Jenkins, 2006) The provision of any further, more detailed documentation is likely to be a matter of what material a

depositor may choose to provide. Most repositories want to encourage depositors, and thus are unlikely to make demands which might deter this, such as requiring resources to be fully documented. (Woolard, 2007) Thus there is already significant variation in local practice in the application of metadata, and this may lead to differences in manner and level of documentation provided for digital resources, (as we discuss with reference to the AHDS below)

LAIRAH Research on Documentation

We have shown why technical documentation is important, and used examples from UCIS research, which attempted to reuse digital text from the OTA. We now employ evidence from the LAIRAH project about user perceptions of procedural documentation, or the lack of it, in digital humanities resources. We also report on the state of documentation provision from the LAIRAH study of resource creators.

Documentation and Resource Creators

As part of the LAIRAH project we interviewed the creators of a sample of well-used projects in digital humanities, to determine whether there were common elements of good practice which might help account for their success. We chose the sample as a result of log analysis of the Humbul Humanities Hub and the AHDS website. We also asked for the expert opinion of AHDS subject centre staff about which projects they considered to be most popular (Warwick et al. 2007) The projects chosen covered a large number of humanities disciplines, and contained a variety of material, including numerical data and images. These projects were as follows:

- Old Bailey online[7]
- Andre Gide Editions project[8]
- French Stars Project[9]
- The English Monastic Archives Project[10]

- The Survey of English Usage[11]
- The London College of Fashion Archives[12]
- Excavations at Eynsham Abbey[13]
- Toronto Dictionary of Old English Corpus[14]
- The Ave Valley Project[15]
- The Avant Garde Project[16]
- The DIAMM Project[17]
- The Channel Tunnel Rail Link Archives[18]
- Designing Shakespeare[19]
- Exeter Cathedral Keystones and Carvings[20]
- The Suffrage Banners Project[21]
- The Jeremy Bentham Project[22]
- PARIP[23]
- The Powys Digital History Project[24]
- The Celtic Inscribed Stones Project[25]
- The Imperial War Museum Concise Art Collection[26]
- GIS of the ancient Parishes of England and Wales, 1500-1850[27]

All responses were anonymised and are referred to below by randomised participant numbers (eg P17).

Internal and external documentation

We found that levels of procedural documentation are highly variable and that there were numerous different approaches to the collection of documentation. In this sample of twenty one projects we found that all but one had kept some kind of documentation. Older projects, such as the *Survey of English Usage* tended to document fully, since it has become vital to preserve the project's collective memory over its more than fifty year life. However, in the majority of cases procedural documentation was partial or fragmentary and might consist of internal documents such as emails, the minutes of meetings, planning documents or progress log books. Some documents had also subsequently been lost. Documentation was therefore unlikely to cover all aspects of the project, and might be difficult for anyone not involved in the project to understand. As one interviewee put it:

> Yes well, I mean, they might be too chaotic to, you know I mean we didn't really create those for the public it was just rather for us so that we knew what we were doing but, you know they are not really, don't think they are kind of useful for anybody else. (P21)

It might also require significant effort on a potential user's part actually to locate the information they required in such documents, which are rarely indexed or organised. There is also no certainty that all decisions made may be recorded in such internal discussions. Thus such documentation can function as internal project archives, but would be of limited utility as outward-facing records.

Although almost all the interviewees were aware of the importance of documentation, they had been unable to document as fully as they wished because of lack of time. Since documentation was not part of the project deliverables, nor was it as vital as peer reviewed publications, it tended to be neglected.

> I do remember it was quite fraught latterly because there were […] the publisher's deadlines to meet and so on and ironing out the bugs. It was very much a seat of the pants business really. So it was very much operational really rather than, we didn't have the time […] we were in new territory for us we were so anxious to get the thing done that we didn't really have the leisure or indeed the foresight to plot what we were doing. (P22)

The Principal Investigator of one project also offered an interesting suggestion as to why technical experts seem disinclined to document decisions.

> I think IT people tend to be instinctive operators. I mean if […] you sit along side someone and you know they are kind of rattling across the keyboard doing things and the idea of asking them to explain what they are doing, it's just not, you know it's all operating so quickly across the synapses that to record it is an entirely different operation I suspect. It's not, they are not reflective practitioners they are practitioners and I don't mean to in any sense to down play the importance of what they do, I think in fact the efficiently of what they do is predicated on not thinking a second […] thought about it. (P22)

Projects produced by archivists, archaeologists and linguists were documented most fully. Documenting decisions is usual in such disciplines, and therefore producing documentation for a digital resource is regarded as a normal component of the project.

> Yes well that's the sort of scientific paradigm, in a sense that if you are given a pile of Roman pottery then saying what you are doing while you are doing with it and documenting it is seen as part of the, you know, the rigour of the study. (P20)

> […] as long as we put a decent amount of explanation, by the data in this form, this logical train of thought that brought you to that point, in the ADS package, the data package, along with the database, along with all the plans, then it will actually be useable, we hope, by this people who use it. […] perhaps the way best way is to look, if you look at the project on the ADS website, the whole preamble, the text files at the beginning, explain why, what is included, is included. (P13)

Users and Documentation

In this section we discuss the way that users perceive digital resources, whether they were able to find documentation, and what kind of information they thought was needed for digital resources to be useful to them. This documentation might be either technical or procedural. Our participants did not make such distinctions, but were simply interested in finding the right information to reassure them about resource quality. Our findings demonstrate the extent to which projects were able to explain what they had done to their users, whether experts in the resource's discipline, users with more technical interests, or members of the wider interested public.

We held two workshops where users recorded their views of a sample of digital resources, without knowing which were neglected and which had been accessed. (Warwick, et al. 2006) Accessed projects were those for which the log data showed repeated visits through Humbul or the AHDS, or one for which AHDS service providers had multiple requests for access (these measures could of course overlap). Neglected projects were those where there was little or no evidence of access. We did not initially tell participants which projects were in which category, as we did not want participants to assume projects were of poor quality simply because they were neglected.

Participants at the first workshop were academics and digital humanities professionals, and at the other Masters students from UCL SLAIS. Participants noted their impressions of the resources on printed pro-formas, and later discussed their views in plenary session. Responses have, once again, been anonymised, WP followed by a number indicates a written response from the first workshop, and SWP indicates a written response from workshop two. W1 and W2 indicate quotations from discussion at the respective workshop (it was not possible to identify individual participants from spoken comment transcriptions.)

The projects chosen were as follows:

Neglected projects:

- Art and Industry in the Eighteenth Century[28]
- Collected Poems of Wilfred Owen[29]
- Correlates of War Project : International and Civil War Data, 1816-1992 [30]
- Exeter Cathedral Keystones and Carvings [31]
- Other Educated Persons[32]

Accessed projects:

- GIS of the ancient Parishes of England and Wales, 1500-1850

- Imperial War Museum concise art collection
- Toronto Dictionary of Old English Corpus
- Channel Tunnel Rail Link Archive
- Designing Shakespeare
- English Monastic Archives

We also compared the results of the two workshops to determine whether the training that the students had received in the creation and evaluation of digital resources might cause different responses to the sample of projects, but found very few differences, when participants were discussing issues related to documentation.

The Purpose of Resources

Participants at both workshops were often unable to find information about the resources concerning issues which ought to be covered by documentation. They were frequently confused about the purpose and possible uses of resources that were new to them. The three participants quoted below, for example, were giving reasons for their assessment of the quality of the resource, and it is notable that they responded by equating quality with how easy it was to use the resource or understand its purpose. These comments were indicative of reactions, but other similar comments were made repeatedly.

> I don’t really understand what it is. I don’t feel like I know enough to be able to rate it. Could do with some introduction that describes what it actually is and gives some background. I just found it confusing (WP5)

> It looked good, but when I came to use it I didn't know what I was looking at or what to do, so kept going to its parent site. I got lost! What is a GIS?(SWP1)

> I didn't really know where to start here! I didn’t really understand what to search for. Perhaps it's because it's not a resource that I see as relevant to me/my work. (SWP3)

Participants repeatedly commented that some resources seemed to have been designed for the expert user, and tended to deter the majority as a result.

> I think what's important with a lot of these sites is that there is nothing telling you how to approach it. Everyone looks for something in a different way and when someone puts together a resource or whatever it is they come from a particular perspective when they are doing it and you need other people to know what that is. You can't just say "I am a historian, you are a historian so you can use this." You need to say "Well this is where I come from, this is why I did this. This is some information about this." So that someone else can come up with their orientation to the site. Yes I think if you are going to put a site together or any sort of information together you should treat the people that are using it as complete novices. (W2)

They requested more information for non-expert users and argued that simple explanation of the site's purpose and guidance on how to use it is not detrimental to the expert academic or information professional, but is very helpful to the novice user.

Some participants blamed themselves if they could not understand how to use a resource, speculating that perhaps they were not the right kind of user or that they had been using the resource wrongly. The following comments were made about the same resource:

> I didn't understand the 'search text' page- what does 'context- (n) characters' mean? Or 'show matching region'? This site must be designed for web creators? Technical people? Not for people who want to search the [name deleted] poems! (SWP8)

> I don't understand this one. All I can get is the full text of [name]'s poems (which I would rather read from a book) or a way to search phrases/words from his poems. I don't know what I would use this for. Perhaps I'm not using it properly (SWP3)

It is perhaps significant that these two respondents are students who have studied the creation and evaluation of digital resources, yet still feel a lack of confidence in using this new resource. This kind of comment was rare amongst the academics and computing professionals, who, if they found difficulties with using a resource tended quickly to dismiss it as poor. This may reflect different levels of confidence in participants' academic judgment.

This lack of understanding is not simply regrettable, however, as many of the participants made clear that if they could not understand the purpose of the resource, they would assume that it was of poor quality, and as a result would tend not to use it.

Conversely the reasons given for thinking that a project was of good quality often included the fact that information was provided, as these two written comments on the same site indicate:

> Site list easy to navigate-data intuitively placed on screen. Useful info on various historical periods. Good overviews. (WP4)

> Clear, simple; different levels; comprehensive introductory descriptions of texts and resources.(WP8)

This suggests a link between lack of understanding and lack of use on the part of our participants. It also suggests that if creators wish their resources to be used, and considered to be of good quality, it is important that documentation about its purpose and help with use should be provided.

This is a new demand for academics used to writing monographs whose expected readership is likely to be other academics and perhaps research students who already have at least a reasonable knowledge of the area. Thus authors do not need to explain the importance of their subject, and may resist doing so, in case their book is regarded as insufficiently scholarly. In the case of digital resources, the intended audience may be interested amateurs, but are as likely to be academic experts, or technological experts, but not what Holscher and Strube (2000) have called double experts: those with expertise in the use of digital resources and the subject domain. As Holscher and Strube show, only double experts are likely to be truly adept at using digital resources, and thus information on the resource's contents and methods of use is helpful for both single experts and novices.

Content and Provenance

Several participants raised doubts about the quality of the content and its extent.

> It's hard to tell if you are looking at a complete resource by access via searching. (WP13)

One project was still incomplete, and although some participants questioned how reliable searches of it might be, this information about the state of the data was generally welcomed.

Participants found that most of the resources were deficient in the kind of scholarly information about the provenance and selection of sources usually provided in print by citations and bibliographies.

> Not clear how comprehensive or up to date the record is (although this information is available, it needs some seeking out). (WP12)

In contrast, another resource was praised because

> Extensive references are provided which gives the site authority and reassures the user that it is reliable. (SWP4)

Finding Documentation

Participants felt that they needed more information about the resource's purpose and the reliability, provenance, and completeness of its contents, all of which could have been provided by good procedural documentation. However, many participants were unable to find such documentation. Some of the resources did not provide such information, as had been in the case of the OTA data that UCIS tried to use. In some cases documentation was provided, but some participants could not find it. We also found disagreement about the same resource; one participant might insist that no information was to be found, while another thought there was too much, as the following comments about the same resource show:

> It is an up to date resource with lots of information given as to what the resource contains. The description is very plain and nothing about it sells the resource. (SWP2)

> It is good that it covers a variety of sources from different disciplines. Perhaps it could be clearer about how comprehensive it is. (SWP3)

> In order to gain entry to the resource users need to create a data usage profile which isn't immediately clear. Perhaps a list of options could be provided otherwise this could be a

> barrier/deterrent to usage. It is not immediately clear how the data could be utilised either. (SWP6)

> On face value looks very specialised, to the extent that even a specialist would think it someone else's speciality. A lot of the metadata is search metadata; while you should be able to look 'under the hood' none but a librarian/info professional would be anything but put off by HASSET/LCSH.[33] Internet is a combined image/text medium- give photos for context. (SWP10)

The problem is not simply caused by a lack of documentation- the information provided is extensive and complex- but whether users are able to comprehend it. The comments that SWP10 makes are instructive in this context. It appears that a further issue in the provision of documentation is what kind should be provided and to what level of complexity.

When reading on the web, participants take in about 40% less information that reading in print, tend to scan rather than read in detail and dislike scrolling pages (Morkes and Nielson, 1997). Several of the resources that provided documentation tended to provide it either in complex technical language or on long scrolling pages. In one case documentation was distributed across different parts of the website, requiring the user to be relatively determined to track down all the information available. In another case the information was presented on a very long scrolling page, without subheadings or thematic organization to guide the reader. Another project provided very detailed information about its production and provenance, but the information was structured as if designed for print publication; thus the pages were long, and the text was complex, and even included footnotes, presented as if on the printed page.

It was not surprising, therefore, that participants found them difficult to understand. A project with a large amount of metadata was described as ‘dense and intimidating’. (WP8) Another participant commented:

> The resource provides lots of information that is not of any real use to the user such as file structure and no of variables per record. By getting rid of this information the resource would be less cluttered (SWP2).

The same participant was aware that large amounts of text were problematic.

> Too much text/long pages makes it hard for the user to see the benefits of the resource. (SWP2)

An archaeological resource included a large amount of introductory text, and one participant noted that:

> I know a little about archaeology. If I did not some basic explanation would be useful within a few clicks, .eg. major engineering projects-urgency of findings for archaeological recordings. Otherwise there is no link to the provenance of the data in terms of methodology (in layman's terms).(WP7)

This information is within one click of the opening page, but as another participant commented, the introductory page is long, and it may be that the information was not immediately evident. Thus it is possible that some participants simply failed to find the information they needed, although it was provided, because the form in which it was presented was too complex, detailed or difficult to find.

Access to Documentation

We have shown that from the point of view of users, it is important to make documentation as easy to find as possible. However the LAIRAH research showed that even when considering a sample of well used projects, access to documentation could be problematic. As a result of our interviews, we found that if decisions were documented in an informal way, the resulting documents, either paper or electronic, tended to be kept by the individual PI or their institution in a way that was not advertised, and thus evidently not accessible to future researchers. One digital humanities research centre has, from its foundation, recommended that all documentation about projects was to be deposited with the library. This is an excellent way of preserving such information. However a potential user would have to visit the University if they wished to consult documentation, and the facility to do this is not widely publicised.

Ideally, users should be able to access documentation electronically, if possible through the project website itself. However once again there is a wide variety of practice in our sample. Two projects included no documentation at all that we were able to find on their website and two others had no independent website. In contrast

the Powys Digital History Project and the Old Bailey Online made documentation available from their websites, linked from the home page main menu and titled ‘about the project’. The information is then clearly presented in simple terms that non experts could understand, but including enough technical and scholarly detail to satisfy more expert users, and those who might wish to construct a similar resource. It is also well suited to web delivery, given that the information is divided into sections and organised under subject headings such as ‘Project specification’ and ‘Design elements.’

Take in Figure 1.

Other projects provided some contextual information on their websites; however levels of detail varied from the highly detailed and technically complex to the very basic.

Help for Users and Information for Re-use

It is important to return to the dual function of documentation, that is to provide information to aid the user in making best use of the resource, and to provide the kind of technical details that can aid re-use, for those who wish to construct a similar project. On their websites, most projects concentrated more on guiding the user than on information for re-use. Although some projects did include good technical information, in the case of eleven projects out of our sample of 21, this was basic or non-existent.

Technical Documentation via the AHDS

We have shown that the most common standard for documenting information in repositories is likely to be Dublin Core metadata. It is impossible to know how well this will work for digital humanities materials in future; however problems inherent in applying this kind of metadata are exemplified by the practice that the AHDS began to adopt relatively recently. In 2006 the AHDS adopted a central cross search catalogue (shown in figure 2 below) which presented a common set of structured metadata fields to the user if the link from the search results is clicked.

Take in Figure 2.

This means that searchers could access at least basic documentation, and it was also possible to link from this to more detailed documentation, if it had been deposited.

Fields were structured in a similar spirit to that of the Dublin Core metadata set, in that they were composed of a limited set of simple terms, eg title, description, date, rights etc although some field such as Temporal Coverage, and Associated Publications were unique to the AHDS, and some fields were not included in a description if information about, for example rights was not relevant to a particular resource. Nevertheless the tension between structured fields and free text content was still evident. The interpretation of what the fields should contain varied widely; this may have been as a result of the data supplied by the depositor, but also seemed to be somewhat influenced by the service provider. Thus the documentation of resources in Visual Arts tended to resemble the kind of interpretation of images often found in exhibition catalogues, and was relatively lacking in technical information.

If we compare specific examples, the Format field could vary between the simple ‘image database’ to the more complex ‘Markup: HTML 4.0; Markup: Rich Text Format; Binary text: PDF; Database: ASCII comma separated values; Image: AutoCad R14 DXF; Image: PDF; Image: SVG’. The latter would be much more useful for anyone wishing to create a similar resource. The ‘description’ field might include a few sentences of skeleton description of the content, or several paragraphs of highly detailed material, including the provenance of sources. Again these would be of very different utility to a potential user, wishing to understand the extent of the collection. In general the provision of technical information in the cross search results is basic; in some cases the data in the format field is as basic as “DVD” or “TIFF” files. This is unhelpful for technical experts and more general users alike. While most non experts would understand what a DVD is, few would probably know that a TIFF is an image format. Those who did know about TIFFs would almost certainly need more information about, the version of TIFF used, the capture resolution and bit depth at which the images had been scanned, and whether any compression had been used to store the image. Although for some resources, it is then possible to access very detailed information at the next level or two levels down, this is by no means common. Thus in many cases it would be difficult for a potential re-user, or someone

wishing to design a similar project to access sufficient technical information from the documentation provided to understand the rationale for the resource's creation.

Discussion

Technical Documentation

We have shown that even within the field of digital humanities projects there is a great variety of practice in the area of documentation. In terms of its dual function, most documentation available tends to be procedural, and thus more helpful in guiding users of the resource than providing the kind of detailed technical documentation that might help those who would like to learn more about how the resource was created, perhaps with a view to producing something similar themselves, or being able to trust the resource for use in research. Our experience with the UCIS project showed how difficult it is to re-use digital data, and this was caused not only by idiosyncratic markup, but also by a lack of documentation which might have helped us to reconstruct the data creator's practices. We have found from an examination of project websites that it can also be very difficult to access detailed documentation of technical aspects such as database fields, markup schemes or even number and type of digital files. Given the lack of technical documentation, the re-use of digital data may prove difficult.

Production of Documentation

Although projects realised that they ought to keep documentation, interviews with producers showed that it was accorded a low priority because it was not a project deliverable. As P22 remarked

> ..we were so anxious to get the thing done that we didn't really have the leisure or indeed the foresight to plot what we were doing.

Thus only in disciplines where their academic peers would expect documentation as part of a scholarly project did we find it routinely kept. It is perhaps significant that

one of the most effective projects at keeping and presenting documentation in a simple, comprehensible fashion was compelled to do so by its original grant from the New Opportunities Fund[34]. As a result the LAIRAH project report to the AHRC has recommended that the production of documentation should be a deliverable of all research grants where a digital resource is produced. (Warwick et al. 2007)

Access to Documentation

Consistent presentation

Our examination of the AHDS search interface shows that even where particular fields are used to structure documentation and metadata, the results can be of variable quality and complexity. This problem has been noted in the NHS (discussed above) and is also likely to affect institutional repositories. The more variation that exists the harder it is likely to be to cross search repositories for digital humanities material. Despite the confidence that the AHRC obviously has in such repositories, consistency in presenting information is more easily achieved by national bodies, such as INTUTE[35] (formerly the Resource Discovery Network (RDN)[36]). Although subject centres are physically distributed, INTUTE has agreed on uniform presentation of data, and consistent web design, which helps to avoid users becoming disorientated, should they need to move from, for example, INTUTE Arts and Humanities, to INTUTE Social sciences. The study that Palmer and Neumann carried out on humanities researchers suggests that a significant proportion of users may need to do this when engaged in interdisciplinary research. (Palmer and Neumann, 2002)

To be most useful for users, documentation should also be easily accessible via the same web interface as the digital resource itself. It is also important that it be clearly labelled and easily identified- “About the X project”, for example. This means that users do not have to search the entire site for different pieces of information. Just as the ‘contact us’ link has become an expected top level link on web pages, so it would be advisable for digital research projects to have a similar ‘about the project’ link on their home page (Nielson and Tahir, 2001). The Old Bailey project was the most successful project in our sample in this respect as can be seen in the example below.

Take in Figure. 3

Although nine top level links may be too many, there is a clear link to documentation from the project home page. The use of this top level link, although very easily implemented, would be of great help to users.

Our work with users also shows that the level and type of documentation is important. It must be presented in a way that makes best use of web functionality; under thematic headings, and on short pages to avoid scrolling. It also should not be so detailed or complex as to be confusing to most users, while at the same time being complex enough to satisfy the needs of those who need to re-use the data. This may sound an impossible task, but has been achieved in the case of Old Bailey online, where the top level documentation gives information for the majority of new users, then provides links to more detailed material. This would seem to be a promising model to follow. It is also significant that the Old Bailey online is a very well used project. As the interview with its PI demonstrated, its own project logs show very high levels of access. It was also one of the very few non-commercial resources consistently named by respondents to both the LAIRAH questionnaire and that of the Institute of Historical Research's ICT Strategy project. (IHR, 2007) Thus its good design and comprehensive documentation have obviously helped it to achieve a reputation amongst users as a trustworthy research resource.

Take in Figure 4.

Different Documentation Practices

The variation we discovered in documentation practices is partly the result of clashes between digital resources and traditional research cultures in the humanities. Traditionally it did not matter that archaeologists had a different attitude to documenting their findings from art historians or literary scholars. The artefacts that they researched were accessed in different ways, dug from the ground, collected in museums and galleries or printed in books. Their findings were also published in different journals or monographs. Thus each community has formed its own type of practice and research norms in both the analogue and digital research environments,

and it is difficult to abandon such practices simply because technology is changing. (Wenger, 1999)

However, the effect of the web means that users are now accustomed to accessing remote digital data in a variety of forms, from central information services, such as the university library, and search engines like Google. Users wish to find appropriate journals papers for their work, but have little interest in, or awareness of, the contents of the rest of the volume in which the article was originally published. (Mahoui and Cunningham, 2001) In a similar way, users of digital resources may increasingly become less interested in the disciplinary origins of a research project, or the location of an institutional repository, instead simply seeking the appropriate data to further their research. Web delivery of such data must therefore concern itself with the utility of such data for a range of users from different subject backgrounds and of a variety of levels of expertise. Good documentation is vital in this regard, as is some kind of agreement on its presentation.

However it will be much more difficult to solve the problem of the variation of free text data in delimited fields. Even if there can be agreement in the humanities research community about which fields or subjects should be included in documentation, the data in them is likely to differ according to different subject communities, or according to the practice of individual repositories. Research in the area of medicine, discussed above, suggests that this is extremely difficult to avoid, and can only be countered by the kind of intensive monitoring and training that is vital in health records, but is unlikely to be accorded such a high priority in the case of digital humanities research projects.

Conclusions

It is to be hoped that simply by drawing attention to some of the problems that may occur in reuse, our work will cause resource creators to take seriously the importance of documentation and consistency. Our research has examined the issue from the point of view of the producers of digital resources, the users of such resources, and

our own attempt to reuse digital data which was inconsistently marked up and undocumented. The latter is a case study of our experience of one UK-based repository. But since the OTA is one of the most reputable sources of good quality electronic text in the world, our findings should be of interest to the creators and users of other electronic texts well beyond this particular example. Not all electronic texts are of such high quality, nor are they always collected by an archive, and so such considerations become even more important when texts are made available by repositories managed by institutions such as libraries or university departments or even the web pages of individual scholars. It is also clear that users of digital resources already have very high expectations of their quality and that of the accompanying documentation, and this should be applicable to many other digital resources produced by scholarly research projects, not only in the humanities.

Yet our attempt to reuse digital humanities data has shown that lack of consistency and documentation may render this task almost impossible. The advantage of markup languages such as XML should be that data is easily portable and reusable irrespective of the platform on which it is used. Yet the idiosyncratic uses of markup that we found have almost negated this advantage.

The creators of the resources probably thought only of their own needs as researchers and were happy with markup that made sense to them. It is still common for projects that use TEI to create their own customisations, without necessarily documenting them. Unlike most scientists, whose collaborative research practices make them aware of the importance of adhering to standards and conventions that make their code comprehensible, humanities scholars are rewarded for originality, and tend to work alone. Research paradigms do not oblige scholars to think about how their work might be reused, their data tested, or their resource used to further research collaboration. Some projects had good intentions about documenting their work, but have found it difficult to carry this out adequately in the time available. This might not matter if the creators of a resource are its only users, but given the intellectual and monetary cost of resource creation, their authors ought at least to be aware of the possible implications of applying idiosyncratic markup without comments or not providing good documentation. One solution to this problem may be to require documentation as a condition of the grant. Another is to make producers aware of how much users

need this information, and indeed the negative conclusions they are prone to draw about the resource's quality and utility if they are unable to find it. This paper provides the evidence of just such consequences.

Acknowledgements

The UCIS project is funded by the EPSRC grant number GR/S84798. LAIRAH was funded by the AHRC ICT Strategy scheme. We would like to thank all those who took part in workshops, and agreed to be interviewed during the project.

References

AHDS History (no date) Guidelines for Documenting Data. Available from http://hds.essex.ac.uk/docguide.asp

Beacham, R, Denard, H, and Niccolucci, F, (2006) "An Introduction to the London Charter", The E-volution of ICTechnology in Cultural Heritage, Papers from the Joint Event CIPA/VAST/EG/EuroMed Event, 2006, available from http://www.londoncharter.org/Beacham-Denard-Niccolucci_paper.doc

Butler, T, Fisher, S, Hockey, S, Coulombe, G, Clements, P, Brown, S, Grundy, I, Carter, K, Harvey, K, and Wood, J, (2000) "Can a Team Tag Consistently? Experiences on the Orlando Project", Markup Languages Theory and Practice, Vol. 2, No. 2, pp. 111 - 125.

Daneker, I, and Warwick, C, (2005) "A la Carte Schema: A Case Study Comparison of the Application of DTDs and XML Schema to the Carte Calendar Project Template", Peter Liddell, et al.(eds) The Association for Computers and the Humanities-Association for Literary and Linguistic Computing, conference, University of Victoria, Canada June 15-18, 2005. pp. 50-53

Dempsey, L, & Weibel, S L, (1996) “The Warwick Metadata workshop: a framework for the deployment of resource description” D- Lib Magazine, July/August, available from: http://www.ukoln.ac.uk/dlib/dlib/july96/0 7weibel.html.

Dunning, A, (2007) “Re: Documentation”, E-mail to Claire Warwick, 28 February, 2007

Giordano, R, (1995) “The TEI Header and the Documentation of Electronic Texts” Computers and the Humanities, Vol. 29, No. 1, pp. 75-84.

Frackoviak, M, (1999) “Editing and TACT”, available from http://chss.montclair.edu/~sotillos/Tactpages/editing.html

Heery, R, (1996a) “Review of Metadata Formats”, Program: Automated Library and Information Systems, Vol. 30, No. 4, pp. 345-373.

Hockey, S, (2004) “The History of Humanities Computing”, in Ray Siemens, Susan Schreibman and John Unsworth (eds.) Blackwell Companion to Digital Humanities, Oxford, Blackwell, pp. 3-19.

Holscher, C, & Strube, G, (2000) “Web search behaviour of Internet Experts and Newbies”, Computer Networks, Vol. 33, pp.337-346.

IHR (2007) Peer review and evaluation of digital resources in the arts and humanities, Arts and Humanities Research Council. Available from http://www.history.ac.uk/digit/peer/Peer_review_report2006.pdf

Kipp, M, & Campbell, D, (2006) “Patterns and Inconsistencies in Collaborative Tagging Practices: An Examination of Tagging Practices”, Proceedings of the Annual General Meeting of the American Society for Information Science and Technology, Austin, Texas, November 3-8, 2006, available from http://eprints.rclis.org/archive/00008315/

Knight, J, (1997), “Making a MARC with Dublin Core”, Ariadne, Vol. 8, available from http://www.ariadne.ac.uk/issue8/marc/intro.html

Lancashire, I, (1996) “Bilingual Dictionaries in an English Renaissance Knowledge Base”, Computers in the Humanities Working Papers, University of Toronto, available from http://www.chass.utoronto.ca/epc/chwp/lancash1/lan1_3.htm

Mahoui, M, and Cunningham, S-J, (2001) “Search Behavior in a Research-Oriented Digital Library”, Research and Advanced Technology for Digital Libraries, Proceedings of 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001. Lecture Notes in Computer Science, Vol. 2163, Berlin, Heidelberg, Springer, pp. 13-24.

Miller, P, (1996) “Metadata for the Masses”, Ariadne, Vol. 5, available from: http://www.ariadne.ac.uk/issue5/metadata-masses/

Morrison, A, Popham, M, and Wikander, K, (no date) Creating and Documenting Electronic Texts: Guide to Good Practice, AHDS publications, available from http://ota.ahds.ac.uk/documents/creating/

Morkes, J**,** and Nielsen, J, (1997) **“**Concise, SCANNABLE, and Objective: How to Write for the Web” *Useit.com*, available from http://www.useit.com/papers/webwriting/writing.html

NHS Executive (1998) Information for Health: An Information Strategy for the Modern NHS 1998–2005. A National Strategy for Local Implementation, London, Department of Health.

Nielsen, J, and Tahir, M, (2001) Homepage Usability: 50 Websites Deconstructed, Indianapolis, New Riders Publishing.

Orna, E, (2005) Making Knowledge Visible, Aldershot, Gower.

Palmer, C, L, and Neumann, L, (2002) ”The Information Work of Interdisciplinary Humanities Scholars: Exploration and Translation” Library Quarterly Vol. 72, pp. 85-117.

Popham, M, (1998) Oxford Text Archive Collections Policy - Version 1.1, AHDS Publications, available from http://ota.ahds.ac.uk/publications/ID_AHDS-Publications-Collections-Policy.html

Porcheret, M, Hughes, R, Evans, D, Jordan, K, Whitehurst, T, Ogden, H, Croft, P, (2004) “Data Quality of General Practice Electronic Health Records: The Impact of a Program of Assessments, Feedback, and Training” J Am Med Inform Assoc, Vol. 11, pp. 78-86.

Probets, S, and C, Jenkins (2006) "Documentation for institutional repositories", Learned Publishing, Vol. 19, No. 1, pp. 57-71.

Raskin, J, (2005) “Comments are more important than code”, Queue, Vol. 3, No. 2, pp. 64-66.

Sperberg-McQueen, C, M., (1991) ‘Text in the Electronic Age: Textual Study and Text Encoding with Examples from Medieval Texts”, Literary and Linguistic Computing, Vol. 6, No. 1, pp. 32-46.

Sperberg McQueen C, M, and Burnard, L, (2002) “A Gentle Introduction to XML”, TEI P4: Guidelines for Electronic Text Encoding and Interchange, TEI Consortium, available from http://www.tei-c.org/Guidelines2/gentleintro.xml

TEI Consortium (2007) TEI P5: Guidelines for Electronic Text Encoding and Interchange, TEI Consortium, available from http://www.tei-c.org/release/doc/tei-p5-doc/html/

Text Encoding Initiative (2001) Text Encoding Initiative, available from http://www.tei-c.org,

Voss, J. (2007) “Tagging, Folksonomy & Co - Renaissance of Manual Indexing?”, Open Innovation, Proceedings of the 10th International Symposium for Information Science. Constance, UVK pp. 243-254

Walsh, S, H., (2004) “The clinician's perspective on electronic health records and how they can affect patient care”, British Medical Journal, Vol. 328, pp. 1184-1187 .

Warwick, C, Blandford, A, Buchanan, G, & Rimmer, J, (2005) “User Centred Interactive Search in the Humanities”, Proceedings of 5th ACM/IEEE-CS joint conference on Digital libraries, New York, ACM press, p. 400.

Warwick, C, Terras, M, Galina, I, Huntington, P, Pappa, N, (Forthcoming).“If you build it will they come? The LAIRAH study: quantifying the use of online resources in the Arts and Humanities through statistical analysis of user log data”, Literary and Linguistic Computing

Warwick, C, Terras, M, Huntington, P, Pappa, N, Galina, I, (2007) The LAIRAH Project: Log Analysis of Digital Resources in the Arts and Humanities. Final Report to the Arts and Humanities Research Council, Arts and Humanities Research Council, available from http://www.ahrcict.rdg.ac.uk/activities/strategy_projects/reports/index.htm

Wenger, E, (1999) Communities of Practice: Learning meaning and Identity, Cambridge, CUP.

Witten, I, H, Bainbridge , D, Boddie, S, J., (2001) Greenstone: Open-Source Digital library Software, D-Lib Magazine Vol. 7, No. 10, http://citeseer.ist.psu.edu/witten01greenstone.html

Woolard, M, “Re: Documentation, e-mail to Claire Warwick, 27 February, 2007.

[1] The AHDS was an organisation which archived various different kinds of electronic data produced as a result of humanities research. However, at the time of writing funding has just been withdrawn by the UK’s Arts and Humanities Research

Council and Joint Information Systems Committee. It will not therefore continue to accession new collections of data and it is not known what will happen to existing archives.

[2] XML has now largely replaced SGML in new digital publications, although a large amount of legacy data remains in SGML. The attraction of XML is that it is simpler than SGML and was designed to be delivered via the web.

[3] http://www.uclic.ucl.ac.uk/annb/DLUsability/UCIS/

[4] http://www.ucl.ac.uk/slais/research/circah/lairah/

[5] http://www.tei.org?

[6] http://www.ota.oucs.ox.ac.uk/

[7] http://www.oldbaileyonline.org/

[8] http://www.shef.ac.uk/hri/projects/projectpages/gide.html

[9] http://www.shef.ac.uk/hri/projects/projectpages/frenchstars.html

[10] http://www.ucl.ac.uk/history/english/monasticarchives/

[11] http://www.ucl.ac.uk/english-usage/

[12] http://vads.ahds.ac.uk/collections/LCFCA.html

[13] http://ads.ahds.ac.uk/catalogue/projArch/eynsham_OAU/index.cfm?CFID=370757&CFTOKEN=91009870

[14] http://www.ahds.ac.uk/catalogue/collection.htm?uri=lll-2462-1

[15] http://ads.ahds.ac.uk/catalogue/search/fr.cfm?rcn=AVE_MILLET_BA-1

[16] http://www.arts.ed.ac.uk/europgstudies/rprojects/avant-garde/index.html

[17] http://www.diamm.ac.uk/

[18] http://www.ahds.ac.uk/catalogue/collection.htm?uri=lll-2462-1

[19] http://www.ahds.ac.uk/catalogue/collection.htm?uri=pa-1018-1

[20] http://www.ahds.ac.uk/catalogue/collection.htm?uri=va-ECKC-1

[21] http://vads.ahds.ac.uk//collections/FSB.html

[22] http://www.ucl.ac.uk/Bentham-Project/

[23] http://www.bris.ac.uk/parip/

[24] http://history.powys.org.uk/

[25] http://www.ucl.ac.uk/archaeology/cisp/

[26] http://vads.ahds.ac.uk/collections/IWM.html

[27] http://www.ahds.ac.uk/catalogue/collection.htm?uri=hist-4828-1

[28] http://www.ahds.ac.uk/catalogue/collection.htm?uri=hist-4635-1

[29] http://www.ota.ox.ac.uk/ text 2216

[30] http://www.ahds.ac.uk/catalogue/collection.htm?uri=hist-3441-1

[31] http://www.ahds.ac.uk/catalogue/collection.htm?uri=va-ECKC-1

[32] http://www.ahds.ac.uk/catalogue/collection.htm?uri=va-OEP-1

[33] These are abbreviations for library cataloguing standards. HASSET stands for Humanities and Social Science Electronic Thesaurus and LCSH the Library of Congress Subject Headings.

[34] http://www.nof.org.uk/

[35] http:// www.intute.ac.uk/

[36] http://www.rdn.ac.uk/