Electronic Publishing '01 –
2001 in the Digital Publishing Odyssey
A. Hübler et al. (Eds.)
IOS Press, 2001

Only Connect: A study of the problems caused by platform specificity and researcher isolation in Humanities Computing Projects

Claire Warwick and Celine Carty

University of Sheffield
Department of Information Studies
Regent Court
211 Portobello Street
Sheffield S1 4DP, UK
0114 222 2632
Fax. 0114-2780300
c.warwick@sheffield.ac.uk

Abstract

The paper presents work performed in a collaboration between the department of Information Studies and the Humanities Research Institute at the University of Sheffield. In it we investigate problems caused by the use of platform specific software by humanities computing projects. The methodology for the project was that of case studies focusing on projects in the history of Natural history and science: The Hartlib Project (University of Sheffield) The Darwin Correspondence Project, (University of Cambridge) The Robert Boyle project (Birkbeck College, London), The Mueller Correspondence Project and The Newton Project (Imperial College, London).

All projects had encountered problems with the use of platform specific formats, and had decided to use SGML as a way of providing a platform independent format that would also help to preserve the material independent of hardware platforms. This was influenced either by advice from other scholars, contact with humanities computing centres, or with national advisory bodies. In particular, projects considered adopting the DTD developed by the Text Encoding Initiative. Despite this enthusiasm a number of problems were encountered with the type of material to be encoded. All but one project therefore chose not to adopt it in its original form.

The paper argues that problems were caused by researchers working in isolation from each other. This can cause the replication of problems that others have already worked on, and expertise not being shared with different projects. Documentation was often not kept up to date because those working on projects have no time to do this. The paper contends that there is a role for outside reportage in the collection and presentation of documentation and its dissemination throughout the humanities computing community.

0. Introduction

The history of humanities computing projects is a relatively short one in comparison to many disciplines in the humanities. Even the oldest initiatives date from the early 1970s. Until recently these projects were also few in number, and not overseen by any national bodies or advisory standards. As a result therefore projects often developed in isolation and found solutions to problems that they met as these occurred. In this paper we will discuss

the development of a small group of such projects. We will aim to illustrate some of the problems caused by researcher isolation as the domain grew and developed. This will be done by taking as an example the technical problems posed by platform specificity and attempts to solve this using the international standard of SGML markup.

1. Methodology

The methodology for the project was that of case studies of projects in the history of Natural history and science. Because we wanted to gain rich , qualitative data, representatives form each project were interviewed using a semi-structured approach. (Patton, 1990). (Taylor & Bogdan, 1984). The Hartlib Project (University of Sheffield) formed the original stimulus for the study. It faced particular problems coveting legacy data in platform specific formats, as discussed below. We therefore sought to find out whether any similar problems had afflicted other cognate projects. We chose to investigate The Darwin Correspondence Project, (University of Cambridge) The Robert Boyle project (Birkbeck College, London) and The Mueller Correspondence Project. These projects were chosen not only because of their subject matter, but also because they involved the conversion of legacy data rather than the creation of new information.

The interviewees were:
Hartlib Papers Project: Michael Pidd
Darwin Correspondence Project: Dr Alison Pearn
Boyle Project: Professor Michael Hunter and Dr Charles Littleton
Mueller Correspondence Project Harold Short

2. Descriptions of the projects

2.1 The Hartlib Papers project

The Hartlib papers project. (HPP) was established by the Humanities Research Institute, University of Sheffield in 1988, to create a complete electronic edition including full-text transcriptions and facsimile images of the collected papers of Samuel Hartlib (c.1600-1662). Hartlib's contemporaries considered him to be 'the great intelligencer of Europe' He was half-Prussian and half-English, settled in London and exploited his numerous European contacts to create a vast network for the dissemination of knowledge for the education of all mankind (Paterson, 1991). Letters, manuscripts, requests for information, digests of the latest inventions were summarised, copied and sent out again from Hartlib's office on the most diverse range of subjects (Greengrass, 1994: 80) The Hartlib Papers, made up of over 25,000 folios of manuscripts, have been seen as 'a consciously developed database designed to gather, contain and disseminate information on all subjects' The sheer volume of material in the Hartlib Papers made print publication impossible (Leslie, 1990). Hartlib's and his correspondents' views on the interconnectedness of knowledge were another reason for ensuring that the archive be published in a fully searchable way.

The first edition of the Hartlib Papers was published on two CD-ROMs in 1995. However, platform specific software and formats were used. The transcriptions of the original papers were done in Microsoft Word. The choice of TOPIC, made by Verity Inc., as the search engine was intended to allow for highly sophisticated searching, including the building up of complex subject trees in the final product (Greengrass, 1994). Thus the images also had to be in a compatible format to this, the Verity Image Format (VIF).

However, by the late 1990s problem inherent in such choices were apparent. The project consisted of a large amount of data, 125 megabytes (Mb) in over 5,000 separate files, that was in danger of becoming inaccessible. The CD-ROMs were designed to run on MS-DOS and were therefore not compatible with the Windows-based software that was becoming increasingly widespread by that time. They proved difficult to network and were not millennium compatible. This posed long-term preservation problems for the data and the HRI was aware that they must convert the files to formats that had some longevity and were not platform specific. Hartlib made decisions that they subsequently regretted, but had felt to be the best options at the time the project began, given the advice then available to them. Our research aimed to find out whether other projects in the area had faced such problems, and how they might have coped with and recovered from them.

2.2 The Robert Boyle Project

The Robert Boyle Project is based at Birkbeck College, University of London. The Project is working primarily on the preparation of new print editions of Boyle's complete Works and Correspondence. In October 1997 work, funded by the Wellcome trust, began on the electronic edition of Boyle's work-diaries. These are paperbooks comprising approximately 350,000 words, which Boyle filled throughout his life with a wide range of material related to his various interests in literary, scientific and medical topics. (Littleton & Hunter, 1998).

An electronic edition of Boyle's work-diaries could be seen as a continuation of his work, since Boyle himself had begun to sort them by keywords. Each diary entry will be transcribed, including all emendations undergone by the text during its original writing, arranged in chronological order. There will also be hyperlinks to the apparatus of notes and commentary on the background of people, places or processes, as well as links based on keywords and other indexing methods.

2.3 The Darwin Correspondence Project

The Darwin Correspondence Project is the longest-lived project examined in this study. It was established by Frederick Burkhardt and Sydney Smith in 1974, with the aim of publishing the definitive printed edition of all the known correspondence to and from the naturalist Charles Darwin (1809-1882). Although the Project also has staff in the US, the majority of staff are based at Cambridge University Library (CUL) where the largest single collection of Darwin's letters is held. More than 15,000 letters have been located and the bulk of transcription was completed by the mid-1980s (Pearn, 1999b).

Although the publication of printed books remains the Project's primary purpose, the use of computers throughout its history makes it a particularly interesting case study since the Project's life 'spanned almost the whole history of computing in the humanities' (Pearn, 1999b). Although there are plans to convert all the letters into SGML for electronic publication the pilot project will be the conversion of the calendar, which is a list of the letters. This was chosen because it was already highly structured into fields, was complete in itself and was small enough to be completed in a reasonable period of time. It was also considered less likely to be the source of copyright problems, and could be expected to increase interest in the published volume. New letters are being discovered at a rate of about sixty a year (http://www.lib.cam.ac.uk/Departments/Darwin/intro.html) so it requires frequent updating and is therefore well suited to online publication.

2.4 The Mueller Correspondence Project

This project was established in 1988 with the aim of publishing a scientific biography and selected correspondence of Baron Ferdinand von Mueller (1840-1896). Mueller, a German who moved to Australia in 1847, served as chief Government Botanist for many years, was director of Melbourne's Botanic Gardens and has been called the father of Australian botany. The first of three print volumes of selected correspondence was published recently (Home, 1998). After the publication of the print volumes all 12,000 letters will be made available electronically. (Lucas & Short, 1999). The transcriptions of the letters are in Microsoft Word and include extensive critical apparatus. The electronic edition will also include a chronological list of letters, a thematic index and a name register of all the people referred to in the correspondence. It is likely that there will be both a CD-ROM and an online publication.

3. Results

3.1 Hartlib papers

3.1.1 Platform Specificity

In the early stages of text conversion projects researchers often did not realise the problems that the use of proprietary formats might cause. HPP chose Microsoft Word as a positive step. Sheffield University had just adopted Word institution-wide so it was familiar and easy to use (Paterson, 1991). Word was thought to make future conversion easy, particularly by using rich text format (RTF) (Robinson, 1994). They did not realise that this might render data inaccessible in the future until much later in the project.

In the early 1990s no national bodies existed to advise projects about such choices. However, in 1997 a national body, the Arts and Humanities Data service was set up, and began to advise projects that proprietary practices should only be used as a last resort. (Beagrie and Greenstein, 1998). The alternative to proprietary formats and software is to use standard formats. The standard that The AHDS began actively to promote was SGML, since it allows complex capture and representation of data, but limits users as little as possible. (Haynes et al, 1997:37). The Oxford Text Archive , which became one of the service providers for the AHDS had been using SGML for many years. However, with a tiny staff and budget and no brief to publicise the standard, knowledge about its usefulness had been slow to spread. How then did the projects in our sample make decisions about proprietary formats, and whether to adopt standards?

3.1.2 The choice of SGML

The HRI, to which HPP belongs now uses SGML for all its projects. This is partly because of a commitment to what Michael Pidd, the technical manager, called the 'intellectual truth' of SGML. Other reasons were that a standard such as SGML aided preservation, made it possible to publish to multiple media for a variety of different audiences from the same master and to perform sophisticated searching. Another advantage was that the ability to parse documents using the DTD means that 'the emphasis can shift from the technical side onto the editing side'.

3.1.3 The question of DTDs

SGML is a platform-independent, flexible and 'permissive' standard (Haynes et al, 1997). Robinson, 1994, argues that SGML is therefore particularly suited to scholarly text encoding projects. SGML files could also be considered 'a permanent, or even an archival, medium' However, SGML has a number of drawbacks: the mark-up itself is very verbose and SGML itself is complex. Furthermore, writing and testing a DTD can be a long process.

These disadvantages led some member of the scholarly community to feel a need to create guidelines on how to apply SGML to text encoding projects. The Text Encoding Initiative was formed as a result. (Burnard, 1995). SGML was adopted as the chosen syntax of the TEI because it had recently been made an ISO standard and offered the possibility of platform-independent interchange (Hockey, 1996). The TEI aimed to provide a framework which prevented each new text encoding project from having to develop its own SGML tags and DTD from scratch (Robinson, 1994). The first version of the TEI Guidelines was published in 1994, and the most recent version (TEI P3) was released in 1999 (Sperberg-McQueen & Burnard, 1999).

The HRI, however, had not adopted it for the second edition of the Hartlib Papers. Although the HRI follows the TEI Guidelines when writing DTDs, Michael Pidd felt the TEI DTD often made things more complicated for a humanities computing project. One reason was that, because the DTD was designed to be very wide-ranging, it did include sufficient detail, 'since we happen to focus on very specific archives and very specific documents requiring very specific treatments, the TEI DTD just doesn't cover it'.

The sheer size of the TEI DTD made it 'off-putting' to novices, since it includes a large number of files and can seem unwieldy. One of the TEI editors has even admitted that 'the TEI's desire to exclude no-one has led to a multiplication of distinctions at first sight rather bewildering' (Burnard, 1995). Michael Pidd prefers to use TEI-Lite, the slimmed down version of the dtd, and often advises people to look at the TEI-Lite DTD when they are trying to write their own project DTD.

Although it is usually implied that the adoption of SGML must begin with writing a DTD, Michael Pidd 'fundamentally disagrees' with this approach. He usually advises that people begin by deciding what the document structure and tags are going to be without writing a formal DTD. This avoids the need to constantly revise a DTD as new features come to light in the documents. Once the tags and document structure are in place, writing a DTD will take very little time for an experienced person.

3.2 The Robert Boyle Project

3.2.1 The Choice of SGML

Advice from both Jeremy Black at the Sumerian Text Corpus[1] and Michael Pidd at the HRI helped convince the Robert Boyle Project to choose SGML for the electronic work-diaries. Littleton and Hunter (1998) liked its durability and the potential for sophisticated manipulation of the tagged elements. Its platform independence and the advantages of working with a recognised standard, tried and tested on other projects were also important.

---

[1] The Electronic Text Corpus of Sumerian Literature is based at the University of Oxford (http://www-etcsl.orient.ox.ac.uk). It is preparing a complete electronic corpus of over 400 literary works composed in the Sumerian language in ancient Mesopotamia, accompanied by an English prose translation and bibliographical information. All of the texts are being encoded in SGML and made available on the website.

Although this is a relatively young project, what is essentially 'legacy data' must now be converted into SGML, since the initial transcription of the diaries had begun in Word 97 for Windows. This would have allowed easy transfer to Idealist for the initially planned CD-ROM publication (On The Boyle 1997). The project now plans, however, to web publish the material after possible conversion to XML by the HRI.

Like Robinson, Michael Hunter also saw one of the great advantages of SGML to be its roots in the scholarly approach. Charles Littleton said that 'every day that I continue to work on these files in Word is a day wasted [...]. It makes no sense to tag in Word'. However, they were concerned that the use of proprietary software to tag the SGML might in itself cause problems in case the files become contaminated with any proprietary coding or became hard to extract from the package as a result.

3.2.2 SGML in practice

The decision to use SGML raised a number of other questions, in particular that of the DTD. This highlights some difficulties associated with manuscript encoding projects in particular, and with many digitisation projects in general.

3.2.3 The question of a DTD

Initially Charles Littleton considered developing a DTD for the work-diaries, from scratch. However, he soon realised that this would be difficult and time-consuming since he was working alone and did not have a background in SGML. The Project staff had initially felt that the TEI DTD was unwieldy and were aware of the difficulties that the Sumerian Text Corpus had experienced with it. They also felt that the TEI DTD was not well suited to manuscript projects in general as marginalia can be problematic within the TEI DTD and the overall structure of the TEI DTD is designed primarily to deal with the structural requirements of printed materials.

The OTA, who had advised the project to adopt the TEI DTD offered to help the Project staff overcome any problems they encountered. So they decided to use the TEI DTD. They saw this as an opportunity to 'test TEI and see how it works'. The advantage was that the Project staff would not have to attempt this task alone but can benefit from the expertise of the OTA. However, they recognise that they will need to modify the TEI DTD to deal with the problems identified. Even then, it was felt that the solutions offered by the OTA to the problem of marginalia, for example, 'seem to be stretching TEI a bit'.

The time taken to choose the DTD meant that transcription began before a DTD was in place. Charles Littleton, like Michael Pidd, felt that, although people are generally advised not to do so, he 'wouldn't have got to the know the material as I did if I hadn't worked in this way'.

Charles Littleton pointed out how labour-intensive editing for electronic projects can be: it involves a great deal of rather mechanical work which has to be done by someone with a detailed knowledge of the material and of the subject matter. Although the amount of material in the work-diaries is small compared to some digitisation projects, it has still taken longer than anticipated and involved a steep learning curve for the editor, as much in the techniques needed to mark up the documents as in the theory behind SGML. As is often the case, the editor had no previous experience of working with SGML and is completely self-taught. His task is made more difficult by working alone, which he commented could be 'quite isolating'.

This last point demonstrates another advantage of working with the OTA and the HRI, as it allows the editor to collaborate with other people and helps 'reduce the diversity' of mark-up. Electronic text projects can often develop independently of each other, with

staff on each of them working from scratch to develop a DTD or a mark-up scheme. Although Hunter points out that a single standard is not necessarily universally appropriate, lack of discussion about editorial practices and their underlying rationale 'has left each new editor approaching his or her task to reinvent the wheel' (Hunter, 1995: 298).

3.3 The Darwin Correspondence Project

3.3.1 Platform Specificity

The initial decision to use computers on the Project in the late 1970s was purely to sort the 13,000 paper slips which held the calendar entries. Once the calendar entries had been entered on the museum cataloguing package MUSCAT, it became apparent that the computerised records could be used as a way to publish both the calendar and the letters (Pearn, 1999b). Over the next twenty years, however, technological changes have meant that the files have had to be migrated over five mainframe computers and several text-handling programs. Two typesetting languages and three plain-text editors have been employed.

3.3.2 The Choice of SGML

There were several reasons for considering SGML. The CUL was proposing to replace their computer system: the Project's master files were stored on the CUL's mainframe, a VMS ALPHA cluster, but it was unlikely that a proprietary library system would allow them to continue in this way. The Project had also begun to experience problems with their current typesetting package, which, for example, did not allow the incorporation of scanned images. They needed to move to a new 'integrated solution' which would address these difficulties.

Furthermore, in July 1997 the then HRB wrote to all the projects it was funding stating that any computer data generated as a result of their funding should be offered for deposit to the AHDS. The accompanying letter from the AHDS explained the benefits of such deposit, including long-term preservation. This coincided with a growing awareness in the Project of the potential of the electronic database as a resource in its own right (Pearn, 1999b). In response to the letter two members of the Project staff attended an information day at the OTA, where they were introduced to the potential benefits of SGML. (Pearn, 1999a, b).

The Darwin Project is unusual in that it has used platform-independent formats from the outset. The Project was split between the US and the UK, with each site using different platforms and operating systems and hardware. Different software was also used for various stages of data entry, manipulation and typesetting, 'so a device-independent format was essential' and ASCII was chosen . Staff working on the Project have always come from a humanities rather than a computing background, and so have tended to favour standard formats because they did not have the technical expertise to exploit the latest technological developments and they needed to use formats that offered 'disinterested, intelligible and long-term technical advice and help' (Pearn, 1999b). Dr Pearn called this 'one of the best decisions ever made' and it has influenced the Project's attitude to standard formats ever since. This attitude therefore influenced their enthusiasm for SGML.

Although the master files are in ASCII, the Project uses some proprietary software during the editing process. The ASCII files are transferred into Microsoft Word as 'MS-DOS text only with line breaks' for footnoting because Word is more readable and offers the facility to cut and paste text. However, the editorial staff have noticed some problems with such limited use of propriety software:

> *When we were going from Word 6 to Word 97 we are getting contamination in our files, even though we're saving files as text only with line breaks. We are having some problem with contamination.*

This was another factor that influenced the decision to use SGML.

3.3.3 SGML in practice

The use of SGML offered a flexible route to both hard copy and electronic publication It would permit the Project to update their typesetting package and it also allows for better verification of file structure (Pearn, 1999b). This in turn means that the mark-up can be more sophisticated, so that searching of the master files can also be improved. SGML does not 'close down any other options' as it is easy to convert to other formats and it can be refined over time (Pearn, 1999a). In this way, SGML can be seen as offering the solution to all of the technical problems faced by the Project. It also addresses the issue of future-proofing and this was mentioned in the interview as an important reason for choosing SGML.

The fact that SGML is non-proprietary, well-tested, standard and device-independent makes its adoption 'entirely consistent with the lessons learned from twenty-five years of experience' Not only is it non-proprietary itself, SGML also allows the Project to 'reduce present dependence on non-supported software' (Pearn, 1999b). It is even felt that using a standard that is growing in currency will make it easier to hire staff with some existing experience, avoiding the need to train people in Project-specific mark-up.

3.3.4 The question of a DTD

An initial discussion paper stated that 'it would be desirable for our implementation to conform to the Guidelines of the TEI, to ensure the greatest degree of platform-independence' (Pearn, 1999a). This paper also recognised the need for outside help in writing the DTD and establishing the tags to be used. However, in the interview Dr Pearn explained that although the OTA had offered help, the Project had not yet received any real support and so: 'I spent quite a bit of time trying to see how we could use TEI for what we wanted and I couldn't do it, I couldn't make it fit.' She added that she had initially been 'really surprised' that she could not make TEI work but that she had since come into contact with other projects who had encountered the same difficulties with TEI and had also decided against using it. The Project has therefore begun work on writing its own DTD. However, Dr Pearn raised an interesting point about the difference between projects creating SGML files from scratch to those, like Darwin and Hartlib, that are converting existing files.

> *I think that if you're creating new files then it is important to have a good DTD to begin with and really think about what you're trying to do. But if you've already got files, [...] then perhaps the best thing to do is map what you've already got into SGML, so all accents and stuff like that and basic mark-up and basic fields that will map very easily, and not worry too much about having a DTD.*

3.4 The Mueller Correspondence Project

3.4.1 The Choice of SGML

From the outset the project wanted to use SGML, which would be converted to HTML for web publication. The 'fundamental reason' for this was future-proofing: Harold Short pointed out that this was the underlying principle of the TEI. However, another benefit of XML in particular was that it allowed 'maximum accessibility' in electronic databases. The Mueller project were 'not happy working in a proprietary format'. (Harold Short)

3.4.2 SGML in practice

Like the Hartlib Papers Project, the original transcriptions were done in Microsoft Word. However, they aim to convert all the Word files so that the XML files can become the master files for the Project. Most of the conversion work is done on the RTF, since like Hartlib, they were found to be easier to work with. The difficulty of converting legacy data such as Word files was recognised by Harold Short:

> *Where we are now depends quite a lot on the way things were when they started and when people started they did things entirely in the light of the existing technology or in the light of what they then understood about what they wanted to do.*

This statement sums up the experience of all the projects in this study and describes the main difficulty faced by many humanities computing projects.

3.4.3 The question of a DTD

A special DTD was written for this Project. However, Harold Short emphasised that this was only because the TEI DTD for correspondence was not yet fully developed and that they used TEI as far as they could. Apart from his personal commitment to TEI, he also had an organisational commitment as Chair of the ALLC, 'so almost as a matter of principle I would want to use TEI'. The input of the TEI into the development of XML was given as another factor underlining its importance to the future of humanities computing.

However, Harold Short readily accepts that SGML and TEI can often seem daunting, especially to busy academics with no previous knowledge or experience of humanities computing, who are simply looking for a way to future-proof their work

> *SGML is complex because the whole thing of thinking about documents in that kind of way is complex and I think, if people are starting from scratch, it takes a while to understand the implications of what's going on and then when you consider, in addition to the complexities of understanding that aspect of it, there is also the technical barrier.*

He added that the editors of the TEI often encounter this difficulty when they go to speak to people about using TEI and SGML. However, despite having 'a sort of missionary role', he stressed that 'they don't intend to say or even give you the impression that you can't do good humanities research without using the TEI'.

4. Documentation

The previous section has shown, therefore, that many editing practices emerged over time as a matter of trial and error. For example, Darwin's decisions to avoid proprietary formats were directly opposed to Hartlib's adoption of Word. Decisions to adopt SGML were often a result of informal contact between friends, and more latterly influenced by newly formed national bodies. Would it therefore, have been helpful if more formal documentation had been available, or more structured fora for the sharing of expertise?

Documentation is described as vital to a data resource's viability. (AHDS, 1998). Flanders (2000) argues that 'documentation is arguably the most important part of a humanities computing project's long-term existence 'because it allows a project to maintain consistency and continuity internally as well as communicating externally to the larger community'. She emphasises that this contributes to an institutional memory, which would otherwise be lost when project members leave. The AHDS advises that detailed documentation is especially important in the case of projects where non- standard or proprietary formats are used. (Beagrie and Greenstein, 1998). These arguments are undoubtedly sensible, but their dates show that this knowledge had been hard won over a

period of time, and may not have been realised by many projects early in their life cycle. Indeed new projects may still not see this as an important part of their remit. For example, although the CCH strongly encourages projects to keep documentation, the Mueller Correspondence Project only keeps informal records of meetings and emails, since it is not a funded project.

Only the Darwin project in the present study kept substantial documentation stating the reasons and decision-making processes underlying their choices. (Pearn, 1999a, b). Transcription guidelines relating to layout and mark-up have probably existed since the beginning of the project and computerisation began by sorting out the calendar, so structure has always been crucial. Though ironically Dr Pearn observed that it was hard to be certain about this, when working in such a long-lived project. These guidelines have not remained static, however, and they 'still come across some files that have got out-of-date coding.'

In 1995, Hunter's (of the Boyle project) article had discussed the importance of documenting editorial practices and decisions in that way that a print editor would consider usual. (1995: 278). Both Hunter and Littleton, when interviewed, agreed that more detailed documentation on the whole process of a digitisation project is needed. Hunter regretted that the Boyle Project does not even posses a complete 'oral history', much less a written one. Similarly, there is currently no policy to create and maintain documentation on the electronic work-diaries.

While project staff may be too busy to create and maintain project documentation, Michael Hunter felt that there was a 'role for reportage' by an outside body so that some written record is preserved. This could document decisions and processes involved in various projects, in order to provide an overview or guide to 'current best practice'. They saw this as a crucial means of avoiding the 'reinvention of the wheel' and a way of keeping those working on digitisation projects in touch with each other.

Alison Pearn from the Darwin project recognised that documentation is 'such a difficult thing and yet it's something that would be so useful'. When writing the conference paper on the history of computing in the Project, she found that it was hard to gather all the information she needed:

> *There were things where I thought "I don't really know how that happened, I hope nobody asks me". It's ironic that a project that deals with history and the preservation of history hasn't preserved its own history at all.*

One of the founders of the Project, Frederick Burkhardt had considered writing a history of the Project, though this would focus on the paper publication. However, the Project staff are more concerned about technical issues as documentation will be required as part of the deposit of archives with the OTA.

Communication may be improved where projects are related to larger Humanities Computing Centres. Two projects in this study operated within larger units, the CCH and the HRI. Staff in both centres agreed that they can help to overcome the isolation, which leads to people who set out to use SGML or TEI abandoning it because it is too difficult or time-consuming. Staff at the centres are also able to provide technical advice and to keep up to date with new developments in humanities computing, when specialist subject academics have no time to do so.

5. Conclusion

It is apparent that the specific characteristics of a humanities computing project depends very much on the material being dealt with but much depends on decisions made early on. All of the projects reported experiencing problems of some kind with proprietary formats or software, and all were keen to guard against such difficulties by adopting standard formats.

The primary reason given for the adoption of a platform-independent format such as SGML was the need for 'future-proofing'. One of the motivating factors behind the move to SGML seems to be the work of funding bodies and academic data services in encouraging the deposit of electronic archives. However, a variety of other reasons for the use of SGML were also given, including its flexibility, the ability to produce both hard copy and electronic publication from one master and the fact that SGML is a recognised international standard.

Despite the positive attitude of all the projects towards SGML, they reported a number of difficulties with its implementation, especially on terms of writing or choosing a DTD. Although all the projects wanted to follow the TEI Guidelines, it is telling that only one has chosen to adopt the TEI DTD. This was also influenced by Harold Short's personal commitment to TEI and the ALLC. The limitations of the TEI DTD with relation to manuscript encoding in particular were cited as the reason in most cases. It could perhaps be concluded from this that TEI has not yet achieved its aim of providing a simple, accessible standard for all scholarly encoding projects. This is partly due to the inevitable complexity of SGML but also the difficulty posed by trying to be 'all things to all men'. The specificity of the needs of projects working on manuscripts letters and papers often seemed to require more than the TEI DTD offers. There is however a palpable desire within the humanities computing community to embrace such standards.

Perhaps the most unexpected result of the interviews was the issue of documentation. All projects reported the difficulty of creating and maintaining project documentation. However, all commented on how useful such documentation would be, particularly for the older projects, as it would provide an essential link with the past. Those projects based in institutions that have no centralised humanities computing centre pointed out how isolating working in this field can sometimes be.

It seems that often the most useful information and advice comes from other people working in humanities computing but that contact of this kind can be limited. Conferences go some way towards solving this problem. However, there is a clear need for improved communication between different. It was telling for example that many projects were under the impression that the difficulties they experienced with DTDs were unique to themselves, and were unaware that such difficulties were relatively common. The existence of centres such as the HRI and the CCH go some way towards achieving this, but there is room for much improvement, particularly in for projects which have access to little or no institutional support or expertise. Often interviewees were not aware of the existence of some of the other projects discussed in this report.

It was suggested more than once that the present study would be an extremely useful way to improve the documentation of and communication between similar projects. This was something that had not been foreseen when the study was being planned but has turned out to be one of the most interesting aspects of the research, and one which we hope to be able to continue in future.

6. References

All URLs last visited on 17.04.01

AHDS (1998). *Creating a Viable Scholarly Data Resource.* August 1998. http://www.ahds.ac.uk/deposit/viable.html

Beagrie, N. & Greenstein, D. (1998). *A strategic policy framework for creating and preserving digital collections.* Arts and Humanities Data Service. http://www.ahds.ac.uk/manage/framework.html

Burnard, L. (1995). 'Text Encoding for Information Interchange: an introduction to the Text Encoding Initiative. July 1995'. http://www.hcu.ox.ac.uk/TEI/Papers/J31

Flanders, J. (2000). 'Writing about it: documentation and humanities computing'. *Paper given at ALLC Conference, Glasgow, July 2000.*
http://www2.arts.gla.ac.uk/allcach2k/Programme/session1.html#111

Greengrass, M. (1994). 'The Hartlib Papers Project: an electronic edition of the past for the future'. In: Armstrong, C.J. & Hartley, R.J. (eds.), *Changing patterns of online information: UKOLUG State-of-the-Art Conference 1994*, pp.73-87. Oxford: Learned Information Ltd.

Haynes, D., Streatfield, D., Jowett, T. & Blake, M. (1997). *Responsibility for Digital Archiving and Long Term Access to Digital Data: A JISC/NPO study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials.* London: David Haynes Associates.

Hockey, S. (1996). 'The ACH/ACL/ALLC Text Encoding Initiative: an overview.' (TEI J16).
http://quirk.oucs.ox.ac.uk/TEI/History/CS/teij16.txt

Home, R.W. et al. (eds.) (1998). *Regardfully yours: selected correspondence of Ferdinand von Mueller. Vol 1: 1840-1859.* Bern: Peter Lang.

Hunter, M. (1995). 'How to edit a seventeenth-century manuscript: principles and practice'. *The Seventeenth Century*, 10(2), pp.277-310.

Leslie, M. (1990). 'The Hartlib Papers Project: text retrieval with large datasets'. *Literary and Linguistic Computing*, 5(1), pp.58-69.

Littleton, C. & Hunter, M. (1998). 'Boyle on the Web'. *On The Boyle: a newsletter of work in progress on Robert Boyle* (1627-91), 2, pp.8-10.

Lucas, A.M. & Short, H. (1999). 'Anticipating Technical Advance in Recreating a Scattered Correspondence: the Mueller Correspondence Project'. *Paper given at DRH Conference, September 12-15th 1999, King's College, London.* http://www.kcl.ac.uk/humanities/cch/drhahc/drh/abst152.htm

Paterson, A. (1991). 'Windowing the past: a seventeenth century technological archive and its electronic exploitation'. In: Lucker, J.K. (ed.), *IATUL Proceedings: Proceedings of the 14th Biennial Conference of IATUL, Cambridge, Mass. USA, July 8-12 1991: New Technologies and Information Services - Evolution or Revolution?*, pp.164-168.

Patton, M.Q. (1990). *Qualitative evaluation and research methods.* 2nd ed. London: Sage.

Pearn, A. (1999a). 'Electronic development in the Darwin Correspondence Project: discussion paper'. [Unpublished].

Pearn, A. (1999b). 'The Darwin Correspondence Project: evolution of an electronic resource'. *Paper given at DRH Conference, September 12-15th 1999, King's College, London.*
http://www.kcl.ac.uk/humanities/cch/drhahc/drh/abst13.htm

Robinson, P. (1994). *The transcription of primary textual sources using SGML.* Oxford: Office for Humanities Communication.

Sperburg-McQueen, C. & Burnard, L. (eds.) (1999). *Guidelines for Electronic Text Encoding and Interchange.* (Rev.ed., May 1999).http://www.hcu.ox.ac.uk/TEI/Guidelines/

Taylor, S.J. & Bogdan, R. (1984). *Introduction to Qualitative Research Methods: the search for meanings.* 2nd ed. New York: John Wiley.