elpub2002 Proceedings — J. A. Carvalho et. al. (Eds.) — VWF Berlin, 2002

Excavating a Resource

The Electronic Dissemination of Archaeological Grey Literature Using XML and the TEI

Christiane Meckseper[1] and Claire Warwick[2]

[1] University of Sheffield
[2] University College London

Abstract. This paper looks at the usability of XML and the TEI Lite DTD for the electronic publication of grey literature by commercial archaeological units. Grey literature is a valuable but underused resource within archaeology and XML would be a useful tool for the publication of field reports as it would allow a quicker response time and a rapid dissemination of information within the fast moving and changing environment of commercial archaeology. It would also allow practitioners to selectively download separate sections of field reports which are of particular importance to them and to improve the searchability of reports on the web.

1 Introduction

This paper is based on research undertaken as part of a MSc course in Information Systems at the University of Sheffield in the summer of 2001. It takes the form of a feasibility study, investigating the potential of using XML and the TEI Lite DTD for a very particular field and purpose, which is the electronic publication of grey literature produced as part of the practice of commercial archaeology within England.

XML has been recommended by Gray and Walford (1999) for the description of archaeological data and specifically for archaeological site reports for online use in an article published by Internet Archaeology. The research undertaken for this paper was intended to provide the next step into the investigation of using XML for online publication and the site reports of a commercial unit, Archaeological Research and Consultancy at the University of Sheffield (ARCUS) were used as a

case study. The aim of the case study was to analyse the scope and nature of archaeological reports as produced by a commercial unit and the implications for the subsequent mark-up in XML and online publication of those reports. It was also intended to test various DTDs for their suitability for the description of archaeological information and to investigate some of the attitudes within commercial archaeology towards the electronic publication of grey literature.

2 Archaeological Literature

Archaeological data varies greatly in nature and scope and this paper limits itself to field reports produced by commercial archaeology units as part of the planning process, so-called "grey reports" or "grey literature". This partly reflects the background of the author but it is also based on the very specific nature and circumstances of the literature produced within that field.

The Council of British Archaeology (CBA) defines fieldwork publications as "... any work that serves to record and disseminate information derived from a fieldwork project (including watching briefs, evaluations, excavations, surveys of all kinds, and related artefact and ecofact analysis)." (Jones *et al* 2001).

Archaeology by nature is a destructive process, the physical remains in the ground are being destroyed though their excavation and lifting of material. The written record and publication has therefore always been seen as synonymous to the preservation of the archaeological record. It was thought vital that an archaeological site should be able to be reconstructed and reinterpreted from the record preserved on paper and in the publication. Excavation placed a duty on the archaeologist to undertake a "preservation by record" of their site and also to disseminate this data to the interested public and other scholars in the field or related disciplines.

A definition of grey literature as defined by the Luxembourg Convention in 1997 is literature "produced at all levels of government bodies, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers" (Luzi 2000: 112). This can be extended to reports that are not issued for public sale or widespread distribution (Jones *et al* 2001). Traditionally, grey literature was very difficult to access as it was not publicly distributed but archived with each organisation that produced it.

The Internet is now seen as one of the main places for the distribution of grey literature. New methods of management and diffusion of grey literature have been in the form of e-print archives, starting with Hepnet and Arpanet, electronic journals and virtual libraries. Also, XML is strongly advocated for the use of constructing an architecture for the distribution of grey literature (Jeffery 2000).

3 Commercial Archaeology in England Today

Today in England most archaeology is carried out within the framework of Planning Policy Guidance 16 (PPG 16), which determines that archaeological work needs to be considered as part of the planning process. Planning authorities need to assess the archaeological implications of planning applications before those applications are determined. This is closely related to the "polluter pays" principle which states that anybody who wants to develop a site and in effect destroy any potential archaeology on that site, needs to provide the funds to put mitigation strategies in place in order to "rescue" the archaeology through excavation and recording.

This has moved rescue archaeology firmly into the commercial sphere, because as developers now have to pay for archaeological work to be carried out, it is treated like any other building work and put out to tender. The introduction of PPG 16 has also led to the increased production and accumulation of a large corpus of grey literature by commercial archaeology units. This emphasizes the dilemma archaeological practice has been increasingly faced with. On the one hand, it has become just a small part of the planning process and is governed by the same commercial principles as any other development contractor. On the other, the results it produces are still seen as academic and a basis for research. However, the funds provided usually do not allow more than the implementation of an immediate mitigation strategy for development and the production of a minimal report produced for the developer. The reports contain recommendations on further work for the developers and synthesize the results of excavations and evaluations. As it is expensive to publish, usually only a minimum of the raw supporting data, like drawings of features and finds and specialist analyses, are included within the report. Funding for further analysis and publication often has to be agreed separately.

The material (the raw data and the reports) can be accessed publicly in several ways. Excavation archives and reports have to be lodged with the local Sites and Monuments Record office (SMR) which are part of the local government institutions on a county level. Finds are deposited with the local museum and sites of particular interest are published in local and national journals or even as monographs such as the British Archaeological Reports (BAR).

A lot of archaeological data is now being produced in electronic format, like digitized drawings and photography, databases of archaeological contexts and finds and electronic survey data. This can be deposited and accessed centrally with the Archaeological Data Service (ADS) which is part of the Arts and Humanities Data Service (AHDS). The ADS also run a project called OASIS (Open AccesS to the Index of archaeological investigationS), a collaboration between the ADS, the Archaeological Investigations Project (AIP) of Bournemouth University, the Archaeology Commissions Section of English Heritage, and the National Monuments Record of English Heritage (`http://ads.ahds.ac.uk/ project/oasis/`). The project is specifically aimed at generating a bibliography of all grey reports

produced within commercial archaeology and can be accessed via the ADS catalogue. This bibliography contains pointers to the location of the grey reports in the physical world.

4 The PUNS Survey

It was against this background that the Council for British Archaeology (CBA) carried out a survey of the publication practices within the archaeological profession and into the use and usability of archaeological field reports, called "The PUNS survey" (Jones *et al* 2001).

A large part of this report was dedicated to grey literature and the authors stressed that even though fieldwork publications are one of the most frequently consulted types of archaeological literature, grey reports have a limited audience beyond the contractorial and curatorial domain, reflecting the difficulty in access and lack of awareness of that literature (Jones *et al* 2001). Many archaeologists are dissatisfied with this situation as they know that information of relevance to their work is being produced of which they may be unaware.

The survey also looked into detail at how archaeological field reports were being read and used and they highlighted the very selective and non-linear reading of those reports For example the most read parts of a report were the summary and the conclusions and specialists were often particularly interested in their respective sections of reports. For example a ceramic specialist is keen to read the sections of field reports that deal with the pottery found and they would like to extract that data easily.

The survey highly recommend that, against this background, the Internet does have distinct advantages for the publication of grey literature and should be increasingly utilised for the dissemination of such information. In this context, the survey also makes the distinction between "publication" and "dissemination" of material. Putting grey literature online does not necessarily mean that it is formally published. It may be very important to make online users aware of the "grey" and also highly changeable nature of the resource they are using. It can also be argued that the inherently changeable nature of the Internet and its peripheral role, or what is perceived as its peripheral role in terms of publication, may therefore be ideally suited for the dissemination of grey literature.

5 Electronic Archaeological Publication

Electronic publication in general is not new within archaeology. It grew out of the desire to make archives and the raw archaeological data more accessible as the financial constraints placed upon the publication of commercial projects meant

that short synthesis reports were becoming the main medium of publication. The supporting data, which was expensive to publish, often remained locked away in inaccessible archives or was relegated to the back of reports in the form of microfiches.

For this reason the ADS in York was created initially with the specific aim of being a central repository for digital archives. The ADS also aims to provide high quality research discovery tools, through its on-line search engine ArchSearch, that go beyond the confines of the ADS archive and allow the user to simultaneously search various repositories of archaeological data like the National Monuments Record of Scotland (Richards and Robinson 2001). Local and national archives like the Sites and Monuments Records (SMR) are also increasingly operating within digital framework.

Gaffney and Exon (1999) argue that this move from archaeological publication to data dissemination has caused the nature of data exchange to become active rather than passive and that in future this will enable others actively to read, interpret and publish the past (Gaffney and Exon 1999). This view is echoed by Richards and Robinson who argue that the catalogue of the ADS improves the re-use potential of data and, rather than a final resting place, the ADS is only one "stage in a cycle of information gathering and reuse" (Richards and Robinson 2001).

After several decades of an increasing separation between the primary archaeological data and the published synthesis report, electronic publication is now moving again towards the integration of those two components, seeing them as an integral part of the site archive.

The journal Internet Archaeology has put this approach into practice with its publications of two excavation reports, the Ave Valley Survey (Millet 2000) and Anglian and Anglo-Scandinavian Cottam (Richards 2001). The reports are part of an "integrated archive", linking the report within the journal to the digital archive stored by the ADS. Data from the archive can be retrieved when a query from a database within the publication is made. Again, the argument is made that evidence can be immediately queried and the assumptions upon which conclusions are based can be questioned and assessed thus creating more active users of data (Winters 2001).

The approach of the excavation reports in Internet Archaeology is as yet exceptional. The reality of electronic publication, particularly within the field of commercial archaeology and grey literature still looks very different. Despite several electronic journals and a proliferation of web sites set up by fieldwork units, electronic publication of fieldwork reports of the type familiar in print are few and far between and only tend to be additions to the traditional print format (Jones *et al* 2001). Most reports on the WWW, even on sites of commercial units like the Birmingham University Field Archaeology Unit (`http://www.bufau.bham.ac.uk/`), are still only concerned with research projects.

6 The ARCUS Case Study

ARCUS were chosen as a case study as they represent a "typical" archaeological commercial field unit and their reports would therefore be representative of archaeological grey reports within England. ARCUS have so far produced a body of circa 600 field reports and a sample of 10% was chosen for a detailed document analysis. After the analysis a further 10% of that sample was used to be marked up in XML.

The main concerns affecting the conversion of the ARCUS field reports into XML was that the process and results should be easy to apply and be understandable within the strict financial and temporal framework of commercial archaeology. The aim was also to deliver a process and product that would fit into the current agenda of archaeological publishing and to incorporate current standards and practices of data preservation and archiving within archaeology.

A purely archaeological DTD, called ArchML, was recently constructed by David Schloen of the University of Chicago (Schloen 2001). Schloen proposes an entirely new publication paradigm that depends on the translation of archaeological datasets into hierarchical item-based structures, which will then be accessible from any XML-capable browser. Schloen criticizes the current context of publication of digital archaeological data, arguing that researchers are still faced with many different database structures and can make efficient use of data only on a project-specific level. Widespread adoption of his data model would therefore be necessary in order to implement his XML based paradigm. As this research was undertaken against the background of British rescue archaeology which may not yet be prepared to fundamentally re-design their data structures it was decided that the ARchML DTD would not be suitable.

No further archaeological DTDs were found. English Heritage are currently developing an archaeology-specific XML schema for the description of archaeological data that incorporates existing controlled vocabulary (Lee, pers comm.). However, this is a more recent development and this data was not available last year.

DTDs have been constructed for fields that can be incorporated into archaeology such as the Geography Markup Language (GML) for geographical information (`http://www.opengis.org/techno/specs/00-029/GML.html`). The Historical Event Markup and Linking (HEML) project is developing a set of markup and transformation tools that are useful to historians world-wide (Cover 2000). The Museum Documentation Association (MDA), working in collaboration with CIMI and other organisations has developed an XML DTD based on SPECTRUM, an established museum process and documentation standard (Degenhardt Drenth 2001). This DTD is currently being tested, however, it only contains elements that describe the archiving and transferal of objects within museum collections, rather than the objects themselves, and was therefore not suitable for this project.

As the main object of archaeological publication were archaeological texts, rather than archaeological data, it was finally decided to use a DTD specifically

aimed at the markup and processing of text and the DTD of the Text Encoding Initiative (TEI), in particular the TEI Lite DTD, was chosen.

7 The Markup of Archaeological Reports Using the TEI DTD

The TEI DTD is useful for the markup of archaeological reports in several ways. Excavation reports are highly structured documents and the division tags of the TEI DTD would allow the very detailed markup of this structure. This would facilitate greatly the selective retrieval of separate sections of reports. For example, it would be possible to extract a separate corpus of texts from a collection of archaeological reports that consisted of only all introductions or all summary and conclusions. A ceramic specialist would be able to find and extract all pottery reports from a collection and thus access directly the information relevant to her. This type of markup would address the wish for a more targeted retrieval of separate sections of reports mentioned in the PUNS report (Jones *et al* 2001).

The TEI DTD also allows for a very detailed markup of place names and dates, which by their nature occur in abundance in archaeological reports. Other useful elements include information about organizations and sponsors, which would make it easier to search for work carried out by a particular archaeological unit or contractor. Finally, the TEI DTD also has a method for marking up levels of certainty, which could be used to mark up the certainty of dates as well as levels of archaeological interpretation.

Despite the detailed markup allowed by the TEI DTD it was still felt necessary to describe the archaeological information in more detail in order to increase the searchability of reports and also to integrate the data description with information standards in use within the archaeological profession. As mentioned above, one goal of the markup of the reports was to incorporate controlled vocabularies used for archival purposes into the DTD. The controlled vocabularies used for paper and electronic archiving within archaeology currently use the English Heritage thesauri for, amongst other things, the description of monuments, archaeological objects and building materials (`http://www.rchme.gov.uk/thesaurus/frequentuser.htm`).

The incorporation of these vocabularies was intended to allow detailed searching of the reports according to specific monument types or specialist materials. A body of excavation reports could thus be searched, for example, for all sections discussing Medieval burial grounds or for all specialist reports concerning Iron Age pottery assemblages. Therefore three new elements were created: <monument>, <object> and <material>. The elements have the standard TEI attributes with an additional attribute for "schema". "Schema" in this context means the definition of a controlled vocabulary, similar to the schema attribute defined by the Dublin

Core (Woodley 2001) rather than an XML schema. Now that XML schemas are becoming more widely used, it may be prudent to change the name of this attribute in order to avoid confusion.

The elements were chosen as they directly reflect the English Heritage thesauri for the description of archaeological data and because they embrace a very large variety of terms. The English Heritage National Monument Type Thesaurus endeavours to describe the "buried and built heritage" of Britain (English Heritage 2000) and monuments are defined in a strictly hierarchical type definition ranging from promontory forts via hopscotch courts to disarticulated human remains. The <monument> tag can therefore be used to define almost any archaeological feature. For example: "The <monument schema="NMR Monument Type Thesaurus" type="inhumation"> articulated skeletons </monument> were all aligned E-W and laid out with arms across the pelvis."

The <object> tag is based on the MDA Object Type Thesaurus and the <material> tag is based on the EH Main Building Materials Thesaurus which work in a similar fashion to the Monuments Type Thesaurus. More controlled vocabularies could have been incorporated as further archaeological thesauri exist, for example a thesaurus on maritime place names (`http://www.rchme.gov.uk/thesaurus/mar_place/default.htm`) or a terminology for the description of 20^{th} century defensive structures
(`http://www.rchme.gov.uk/thesaurus/def_brit/default.htm`).
However, for the purpose of this paper the three additional tags were seen as sufficient as the terminology covered by those tags could be used to describe more than adequately the range of archaeological information encountered.

8 Conclusions

This paper has shown that there is a definite need for the electronic publication of archaeological grey reports. Despite the bibliographic efforts of the OASIS project, grey literature is still almost impossible to access for archaeologists as well as the wider public. The PUNS user survey as well as the survey that is part of this paper have shown that especially field archaeologists feel that there is a large body of valuable information in existence but that it is inaccessible.

Within commercial archaeology the same or adjacent sites are often dug by competitive units and consequently the data and report for work undertaken is archived in separate locations. It would therefore be vital to easily access the reports done by other archaeological contractors. This information could be important as part of the background research for a new project but also in order to produce more wide-ranging syntheses of archaeological periods or subjects. The faster response time of electronic publication would also make it easier to keep up with the development of projects on the ground and would make it possible to quickly update reports and

make new information available. It would also allow a more direct communication between the producers and users of archaeological information.

Grey literature has been called an extremely valuable asset as it represents the collected knowledge and know-how of an organisation. Presently this asset is not recognised and considerably under-used. Unfortunately, because commercial archaeology is a highly competitive process, there is often a reluctance to make information widely accessible. On the other hand, several of the archaeologist surveyed acknowledged that the electronic publication of their reports could considerably raise the profile of their organisation and be used for publicity purposes as well as for the dissemination of information.

In general the attitude towards electronic publishing is becoming more and more favourable, with all of the surveyed units expressing a strong desire to publish their reports on the web. Unfortunately attitudes have not yet changed enough in order to create an environment that makes electronic publication of grey reports possible in practice. Especially within commercial archaeology, units cannot operate on their own. It has been recognised that there is a need for better recognition of electronic publication on the side of national institutions like the SMRs and especially development control officers. If the funding of improved publication, let alone electronic publication, was put down as a strict requirement by development control, it would be easier to build those costs into a project and to persuade developers that it was a vital part of their mitigation strategy.

The PUNS survey (Jones *et al* 2001) is also calling for more national funding in order to facilitate nation-wide syntheses of material. In order to raise the profile of electronic grey reports it would also be important to find more centralised or easily accessible places of publication than a unit's own website. Websites still present very dispersed information on the Internet and if a site is not indexed properly, the information stored on it will not be easy to find. The ADS already welcomes digital excavation reports and has begun to use XML for some aspects of their data storage and description (Kenny, pers comm.), however, as yet not many commercial units actually store their data with the ADS. This is due to time and money constraints and data still being in a non-digital format.

The practical mark-up of two ARCUS excavation reports has shown that the TEI Lite DTD could be a very suitable DTD for the conversion of texts into an electronic format. The structural mark-up of the TEI DTD would allow users to retrieve and view selected sections of a report or body of texts, for example all specialist reports or all introductions and conclusions. This would directly answer some of the user needs mentioned in the PUNS survey (Jones *et al* 2001) above. To implement this on a large scale the process of marking up the reports and the markup itself would need to be further refined in order to make the documents into fully compliant TEI XML and to achieve the best results in terms of data retrieval. It would also be of paramount importance to incorporate image files into the published electronic reports. In addition it would very interesting to consider

the possibility of including the archaeological elements into the document in the form of namespaces, as suggested by a member of the audience at the ALLC/ACH conference in Tuebingen in July 2002.

One of the great advantages of electronic publication as opposed to traditional publication that has been repeatedly mentioned in this paper is the quicker response rate from the field to publication and dissemination of data. Data that is quickly and widely disseminated (in the form of raw data as well as syntheses) will allow the re-use of that data and thus the formation of a multitude of interpretations. It would also allow a faster return of new ideas into the field and thus help to fuel the practical excavation process with more advanced theoretical ideas and knowledge gained from the quick analysis of excavated material (Hodder 1999).

Commercial archaeology units need to realise the assets they possess in the form of their grey reports. If the knowledge and information contained in those reports was more widely and publicly disseminated, this exchange of ideas and re-usability of data might feed back into the archaeological process. This should be a way for commercial archaeology to overcome the theoretical and practical dead end it has found itself in within the last decade. Vital archaeological information that is so far being hidden in the vaults of commercial units could also feed back into the research process and allow more comprehensive syntheses to be brought together that would help to advance the field of archaeology as a whole.

References

Cover, R. (2000). *Historical Event Markup and Linking.* The XML Cover Pages. [Online]. Available at: `http://XML.coverpages.org`. [Accessed: 14.9.2002].

Degenthart Drenth, B. (2001). *Building on the mda SPECTRUM-XML DTD for Collections Management Data Interchange.* [Online]. Available at: `http://www.archimuse.com/mw2001/papers/degenhart/degenhart.html` [Accessed: 14.9.2002].

English Heritage (2000). *Introduction.* National Monuments Record Thesauri. [Online]. Available at: `http://www.rchme.gov.uk/thesaurus/newuser.htm`. [Accessed: 14.9.2002].

Gaffney, V. and Exon, S. (1999). *From Order to Chaos: Publication, Synthesis and the Dissemination of Data in a Digital Age.* Internet Archaeology 6. [Online]. Available at: `http://intarch.ac.uk/journal/issue6/gaffney_toc.html` (Accessed: 14.9.2002).

Gray, J. and Walford, K. (1999). *One Good Site Deserves Another: Electronic Publishing in Field Archaeology.* Internet Archaeology 7. [Online]. Available at: `http://intarch.ac.uk/journal/issue7/gray_toc.html`. (Accessed: 14.9.2002).

Hodder, I. (1999a). *Archaeology and global information systems.* Internet Archaeology 6. [Online]. `http://intarch.ac.uk/journal/issue6/hodder_toc.html` (Acessed: 14.9.2002).

Hodder, I. (1999b). *The Archaeological Process: An Introduction.* (Oxford: Blackwells).

Jeffery, K. G. (2000). *An architecture for grey literature in a RandD context.* The International Journal on Grey Literature. 2.2: 64-72.

S. Jones, A. MacSween, S. Jeffrey, R. Morris, and M. Heyworth: (2001). *From the Ground Up. The Publication of Archaeological Projects: a user needs survey.* [Online]. Available at: `http://www.britarch.ac.uk/pubs/puns/index.HTML`. [Accessed: 14.9.2002].

Luzi, D. (2000). *Trends and evolution in the development of grey literature: a review.* The International Journal on Grey Literature. 1.3: 106-116.

M. Millett, F. Queiroga, K. Strutt, J. Taylor and S. Willis (2000). *The Ave Valley, northern Portugal: an archaeological survey of Iron Age and Roman settlement.* Internet Archaeology 9 [Online]. Available at: `http://intarch.ac.uk/journal/issue9/millett_toc.html`. [Accessed: 14.9.'02]

Richards, J. (2001). *Anglian and Anglo-Scandinavian Cottam: linking digital publication and archive.* Internet Archaeology 10. [Online]. Available at: `http://intarch.ac.uk/journal/issue10/richards_toc.html` [Accessed: 14.9.2002].

Richards, J. and Robinson, D. (eds.) (2001). *Digital Archives from Excavation and Fieldwork: Guide to Good Practice Second Edition.* AHDS Guides to Good Practice. [Online]. Available at: `http://ads.ahds.ac.uk/project/goodguides/excavation`. [Accessed: 14.9.2002].

Schloen, D. (2001). *Archaeological Data Models and Web Publication using XML.* Computers and the Humanities, 35: 123-152.

Winters, J. (2001). *Editorial: Integrating the archive.* Internet Archaeology 9. [Online]. Available at: `http://intarch.ac.uk/journal/issue9/editorial.html`. [Accessed: 14.9.2002].

Woodley, M. S. (2001). *Glossary.* Part of: Dublin Core Metadata Initiative. [Online]. Available at: `http://dublincore.org/documents/2001/04/12/usageguide/glossary.shtml#S`. [Accessed: 14.9.2002].