Introduction

This paper takes the form of a feasibility study, investigating the potential of using XML and the Text Encoding Initiative (TEI) Lite Document Type Description (DTD) for a particular field and purpose—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 a paper published by Internet Archaeology. The use of XML is proposed by Gray and Walford as it would allow researchers to search for site reports not only on the basis of bibliographical information, as proposed by the Dublin Core metadata scheme, but also on the basis of the nature and contents of the archaeological information. The paper contains an example of a fictional site marked up in XML.
using a small provisional archaeology-specific DTD. Gray and Walford stress that their research should be seen as work in progress and that more work needs to be undertaken in the form of identification of user needs and design of a suitable DTD.

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 markup 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 vary greatly in nature and scope, ranging from research-based field projects that are conducted over a number of years to short-term commercial projects that can last as little as half a day. The data include excavation reports, desk-based assessments of archaeological potential, reports by specialists on material such as animal bone or copper alloys, scientific laboratory reports on dendrochronological or carbon 14 dating, and reports on geophysical surveys.

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). Such publications are, by definition, publicly available and range from monographs through papers in national, regional, and local society journals, to summary reports in annual gazetteers.

Archaeology is a destructive process: the physical remains in the ground are destroyed through their excavation and lifting of material. The written record and publication have therefore always been seen as synonymous with 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. Excavation placed a duty on the archaeologist to undertake a ‘preservation by record’ of their site and also to disseminate these data to the interested public and other scholars in the field or related disciplines.

In this context the original notion of an excavation report was not
only that it represented a comprehensive archive of an excavation but also that it was possible to distinguish between a presentation of the pure facts within this paper archive, and the interpretation by the respective author and excavator. This notion has since been challenged and has also had a direct influence on the rise of electronic publication within archaeology.

Grey literature is defined by the Luxembourg Convention in 1997 as 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, p. 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 by the organization that produced it.

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

3 Commercial Archaeology in England Today

Most English archaeology is carried out within the framework of Planning Policy Guidance 16 (PPG 16). Planning authorities must assess the archaeological implications of planning applications before they are granted. This is closely related to the ‘polluter pays’ principle, which states that anybody who wants to develop a site and destroy any potential archaeology must fund mitigation strategies to ‘rescue’ the archaeology through excavation and recording.

Mitigation strategies state that archaeological remains should be preserved in situ and, should this not be possible, that developers must make adequate provision for archaeological excavation and recording. Developers now have to pay for archaeological work to be carried out and this work is put out to tender. 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. There is an innate duty within archaeology to disseminate results and to produce syntheses of material. 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 ‘developer’s report’. Funding for further analysis and publication often has to be agreed separately.

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. The reports are produced for developers as a result of the mitigation strategies put into place during the planning process.
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, such as drawings of features and finds, and specialist analyses, are included within the report.

3.1 Access to 'grey' archaeological information
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); SMRs are part of the local government institutions on a county level. In some cases a pro-forma sheet containing summary of the site and location of the archive is deposited with English Heritage, the central government body responsible for the management of all heritage resources within England. Finds are deposited with the local museum, and sites of particular interest are published in local and national journals or as monographs such as the British Archaeological Reports (BAR).

Many archaeological data are now produced in electronic format, such as digitized drawings and photography, databases of archaeological contexts and finds, and electronic survey data. These data can be deposited and accessed centrally with the Archaeology Data Service (ADS), which is part of the Arts and Humanities Data Service (AHDS). The ADS also runs a project called OASIS (Open Access to the Index of archaeological investigations) in collaboration with 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 aims to generate 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 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 feel that information of relevance to their work is being produced of which they may be unaware.

The survey also looked in detail at how archaeological field reports were being read and used, and highlighted the selective and non-linear
reading of them. It found that the most read sections 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 that deal with the pottery found and would like to extract those data easily.

The survey recommends that the Internet has distinct advantages for the publication of grey literature and should be increasingly utilized for the dissemination of such information. However, it also stressed that electronic publication can still be a very controversial field and called for better institutional and government support to make it a more widely recognized and acceptable medium of dissemination and publication.

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. The financial constraints placed upon the publication of commercial projects meant that short synthesis reports were becoming the main medium of publication and the supporting data often remained locked away in inaccessible archives or were relegated to sheets of microfiche at the back of reports.

The excavation report had lost its role as a published data archive and archaeologists had also realized that the distinction between data and interpretation was often not as easy to maintain as previously assumed. Also, during the 1980s Critical Theory began to have an increasing influence on the practice of archaeology. This led to severe criticism of the practice of publishing synthesis reports and the interpretation of archaeological data as the main format of dissemination of archaeological knowledge. According to Shanks and Tilley (1987), synthesis reports represented exercises in 'domination and control' by individuals who tried to impose their view of the past on everyone else. According to Tilley, interpretation should be a reflexive and social activity, producing site reports that 'attempt to capture at least some of the ambiguities, disjunctions and contradictions inherent in various modes of interpretative understanding' (Tilley, 1989, p. 279). It was seen as crucial, not only by critical theorists but also by the wider archaeological community, to find ways to distribute raw data so as to give a wider audience the opportunity to create their own interpretations.

The ADS 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 online search engine ArchSearch, that go beyond the confines of the ADS archive and allow the user to simultaneously search various repositories of archaeological data such as the National Monuments Record of Scotland (Richards and Robinson, 2001). Local and national archives such as the SMRs are also increasingly operating within a digital framework. Richards and Robinson state that in the context of increased publication of only synthesis or summary reports of fieldwork, digital archives are becoming increasingly important.
as they may well become the only source for primary data (Richards and Robinson, 2001).

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 reuse 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. Richards and Robinson (2001) state that

the report should still interpret the site in the light of the original research agenda […] without requiring the user to ‘drill down to the data’. The intention behind making the data available and linking it to the interpretation is so that users can examine the assumptions upon which the interpretations presented in the report rest. Users are facilitated to develop alternative models based on the data, which might also be published as part of an ongoing and dynamic electronic seminar.

The journal *Internet Archaeology* has put this approach into practice with its publications of two excavation reports, the Ave Valley Survey (Millet *et al*., 2000) and Anglo 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. Evidence can immediately be queried and the assumptions upon which conclusions are based can be questioned and assessed, thus creating more active use 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 rare and tend to be additions to the traditional print format (Jones *et al*., 2001). Most reports on the web, even on sites of commercial units such as the Birmingham University Field Archaeology Unit (http://www.bufau.bham.ac.uk/), are still concerned only with research projects.

### 6 ARCUS Case Study

ARCUS was chosen as a case study as it represents a ‘typical’ archaeological commercial field unit and its reports would therefore be representative of archaeological grey reports within England.
Publication of Archaeological Excavation Reports

The main concern 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.

It was clear from the start that no attempt would be made to construct a purely archaeological DTD from scratch. This would involve highly political issues of data description and standardization within archaeology and it was felt that the author was not in a position to undertake such a task. When questioned, Dr Julian Richards from the ADS acknowledged that such a process would take even professional archaeologists several months or years and would involve collaboration with national bodies such as English Heritage and the ADS.

A purely archaeological DTD, called ArchML, was recently constructed by David Schloen of the University of Chicago (Schloen, 2001a). Schloen proposes a 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, 2001b). 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 to implement his XML-based paradigm.

Schloen’s DTD is designed to describe the spatial organization, temporal sequence, and high variability of archaeological data within his data model, against the background of Near Eastern research archaeology. There is increasing dissatisfaction with the current class-based data model within British archaeology. A recent research proposal for new approaches to post-excavation analysis of complex, digitally recorded excavations expressed encouragement for ‘approaches in which the various classes of archaeological data—stratigraphic, architectural, artefactual, ecofactual, sedimentological, etc.—can be better integrated’ (Perring et al., 2001). Our current data models may change in the future, but this research was undertaken against the background of British rescue archaeology (in which even the concept of a database can be alien) and therefore it was decided that the ARchML DTD would not be suitable.

No further archaeological DTDs were found. 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 worldwide (Cover, 2000). The Museum Documentation Association (MDA), working in collaboration with CIMI and other organizations, 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 contains elements that describe only 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 publication was archaeological texts, rather than 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 allows very detailed markup of this structure. This would facilitate 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 introductions or summaries and conclusions. A ceramic specialist would be able to find and extract all pottery reports from a collection and thus access directly the relevant information. 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, albeit in a slightly different form than intended, could be used to mark up levels of archaeological interpretation or certainty of statements.

7.1 Extension of the TEI DTD

Despite the detailed markup allowed by the TEI DTD, it was still necessary to describe the archaeological information in more detail to increase the searchability of reports and 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. English Heritage thesauri are used for paper and electronic archiving within archaeology to describe, amongst other things, 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; for example, sections discussing Medieval burial grounds or 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
Publication of Archaeological Excavation Reports

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 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 works 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 twentieth-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. The last alteration made to the TEI Lite DTD was the addition of a ‘type’ attribute to the <date> element. This allows dates to be distinguished between archaeological period date such as ‘Mesolithic’, ‘Roman’, or ‘Viking’, a century date such as ‘1642’, or a modern rendition of a date.

The markup implemented on a selection of ARCUS excavation reports during the time of the research project consisted of the main TEI textual elements as well as the additional archaeological tags. The tags were added to only the non-technical summaries of the marked-up reports, because this would be the most likely place to search a report for its archaeological contents. It was also necessary to test the tags for their searchability before an entire report was marked up in such detail.

8 Attitudes towards Electronic Publication

To set the case study in context, a small user survey was undertaken by e-mail of several commercial archaeological units in England. The survey intended to collect information about some of the attitudes of com-
commercial units towards electronic publication of their grey literature and about some of the practicalities involved.

The survey made clear that most units are keen to see their reports online but find it difficult to pursue a policy of electronic publication in the current climate of developer attitudes and concerns about funding. The size of a unit might make a difference, as larger units have larger projects with more considerable sums and contingency funds available. Small commercial units operate on a subsistence level and cannot afford anything that goes much beyond providing a service for the client.

Most representatives of the archaeological units also stressed that units cannot 'go it alone', meaning that support for electronic publication must come from development control officers who are in a better position to put pressure on developers to fund adequate recording and publication of their sites. If development control, curators, and archaeological consultants started to take electronic publication on board, the provision of digital formats could be written into the project brief and adequate funding provided. It would also be important for more national and regional archives such as the SMRs to accept archives and reports in digital formats.

There was a consensus that many grey reports were too ephemeral to be published in full but that it might be beneficial to publish an index or a collection of abstracts on the web with pointers to the full report. In addition, the Internet could provide a welcome forum to publish more substantial or 'interesting' reports on excavations in full. The disadvantage of this approach is that there will never be a consensus on what constitutes an 'interesting' report.

The Internet would also be a good place to disseminate reports on evaluations or watching briefs that would not have been published in a traditional format. Probably with an eye on funding, most archaeologists agreed that electronic publication would not drastically change the format of the paper-based report but that in most cases hyperlinks would be included to link the report to related paper and other sources of information.

The archaeologists interviewed also agreed that it was still important to publish sites in paper-based journals and monographs as they are still the accepted medium of publication amongst academics and allow dissemination of information to individuals who have not yet caught up with the electronic revolution.

9 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. Our findings support those of the PUNS survey, and show that 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 are 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 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 is an extremely valuable asset as it represents the collected knowledge of an organization. At present, this asset is unrecognized and considerably underused. Several of the archaeologists surveyed acknowledged that the electronic publication of their reports could considerably raise the profile of their organization 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 desire to publish their reports on the web. Unfortunately, attitudes have not yet changed enough to create an environment that makes electronic publication of grey reports possible in practice. Especially within commercial archaeology, units cannot operate on their own. There is a need for better recognition of electronic publication by national institutions such as the SMRs and especially development control officers. If the funding of improved publication, let alone electronic publication, was required 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) also called for more national funding to facilitate nationwide syntheses of material. To raise the profile of electronic grey reports it would also be important to establish more centralized or easily accessible places of publication than a unit’s own website. Websites still present very dispersed information on the Internet and if they are not indexed properly, information will not be easy to find. The ADS already welcomes digital excavation reports; however, it has become clear from the user survey that as yet few commercial units store their data with the ADS. This is due to time and money constraints and data still being in a non-digital format.

It would be vital to look further at the role of local and national SMRs, as they are the main point of access to archaeological data by practitioners and the wider public. A small number of SMRs operate in a digital format; however, it would be useful if those organizations also stored the electronic version of a final excavation report and SMRs were generally digitized.

The practical markup 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 markup 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). To implement this on a large scale the process of marking up the reports and the markup itself would need to be further refined 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 (http://www.w3.org/TR/REC-xml-names/).

One of the great advantages of electronic publication as opposed to traditional publication is the quicker response rate from the field to publication of data. Data that are quickly and widely disseminated in the form of raw data as well as syntheses could be reused and thus the formation of a multitude of interpretations might be facilitated. New ideas could be returned to the field more quickly. The practical excavation process could be enriched with more advanced theoretical ideas and knowledge gained from the quick analysis of excavated material (Hodder, 1999).

Commercial archaeology units must realize the assets they possess in the form of grey reports. This should be a way for commercial archaeology to overcome the theoretical and practical dead end it has found itself within the last decade. Vital archaeological information that is so far being hidden in the vaults of commercial units could feed back into the research process and allow more comprehensive syntheses to be produced, and this would help to advance the field of archaeology as a whole.

References


Publication of Archaeological Excavation Reports


