The Importance of Metadata to Archaeology: One View from within the Archaeology Data Service

Paul Miller

Abstract

The provision of valuable archaeological data across a medium such as the Internet is all very well, but it is becoming increasingly difficult for potential users of your data to actually find it. The current unstructured indexing of the World Wide Web by search engines such as Alta Vista is insufficient for the needs of the Archaeology Data Service (ADS), but it also appears that many of the more complex indexing schemes available are also unsuitable, largely due to their very complexity. The ADS has undertaken to explore means by which disparate archaeological resources may be coherently yet simply described, and this paper will report upon a number of the issues addressed in developing a suitably encompassing yet streamlined metadata system, before providing a flavour of the manner in which the completed scheme will operate.

1 Introduction

Metadata. This word is increasingly to be seen bandied about by information professionals from a range of disciplines -- including archaeology -- but its meaning remains unclear to many, and the potential benefits are rarely elucidated in such a way as to make sense to the uninitiated (Miller forthcoming).

This paper introduces the concept of metadata for an archaeological audience, and discusses some of the ways in which metadata is -- and might be -- an aid to archaeological endeavour.

An important point to bear in mind with current metadata research is that this international cross-disciplinary effort is proceeding extremely rapidly, such that the content of this paper -- composed whilst sat in the sun on a summer's day in August of 1997 -- runs the risk of being superseded by developments between now and its publication in the -- doubtless wet and horrible -- northern hemisphere Spring of 1998. In an attempt to maintain currency, metadata issues are outlined in general -- and hopefully timeless -- terms, with more detailed issues being referred to by means of pointers to external resources, primarily on the World Wide Web which is itself the main driving force behind much current metadata research.

2 What is metadata?

Metadata is commonly defined simply as 'data about data'. Whilst undeniably correct, this definition is perhaps less helpful than it might be in introducing a potentially confusing topic to readers.

More helpfully, the Archaeology Data Service considers metadata as 'a means of turning *data* into *information* [intelligible and of value to those other than the creator of a resource]'. Metadata, then, may be seen as the extra information associated with any object or resource which allows a viewer to place it in context and make sense of it.

In a traditional paradigm such as paper publication, the metadata for a publication might include the author's name, the title of the publication, the publisher, *etc.* It can be seen, therefore, that the metadata for a book is actually much the same as the cataloguing information that might be recorded for that book by a librarian constructing a library catalogue. Here, as elsewhere, metadata often serves as a generic term for procedures that have been established over many years, rather than being something wholly new. Indeed, some of the initiatives currently thought of as metadata standards have been evolving for many years. The MAchine Readable Cataloging (Library of Congress 1997) system used in libraries all around the world, for example, was originally developed to define the manner in which catalogues should be described for storage on computer tape for transfer between different libraries.

3 Metadata for resource discovery

Metadata is a generic term, used to span all aspects of resource (or data) management and manipulation from discovery to storage and reuse. Each of these -- and other -- areas is extremely important in archaeology and elsewhere, and deserving of greater exploration, but the aspect of metadata attracting greatest interest at present is that of *resource discovery*.

Resource Discovery is the process by which a potential user searches for and locates information of use to them. The process may well also include some means of evaluating the resource's fitness for purpose before it is actually retrieved. With the phenomenal growth of resources on the Internet's World Wide Web, and the need for some means by which those searching on-line might avoid the ever-present danger of information overload (Miller 1996), much of the current research into resource discovery is being directed towards the Web community. Many of the lessons learned on the Web are of equal value off-line and apply as much to searching for books in a library or excavation records in an archive as they do to locating electronic resources in cyberspace.

Whereas many of the best known metadata schemes exist to provide subject-specific descriptions of particular resource types for tightly delimited purposes (MARC to describe the contents of library catalogues or the *Content Standards for Digital Geospatial Metadata* (Federal Geographic Data Committee 1997) to describe digital map data, for example), an important aspiration of those defining the metadata appropriate for resource discovery is to develop a scheme of cross-disciplinary value. It is one thing to develop a means of describing a map, a library book, or an archaeological excavation, and quite another to devise a single system for describing all three. This task is rendered yet more difficult when it is considered that the resulting system must describe resources in a manner which appears intuitive to spatial scientists (looking for a map), librarians (managing their books) and an archaeologist (after the records from an excavation) whilst remaining compact enough as not to prove too daunting to those tasked with entering data into it or retrieving material from it.

4 Dublin Core and the Archaeology Data Service

Given the role of the Archaeology Data Service (Wise and Richards, this volume), and its reliance upon electronic delivery of data, developments in resource discovery are being watched extremely closely (Miller and Wise 1997, Miller and Greenstein 1997), and great hope is being placed in one particular resource discovery initiative; the Dublin Core.

The Dublin Core (Weibel *et al* 1997) arose from an international cross-domain effort to define a means primarily of providing author-generated metadata for small-scale resources on the World Wide Web. This effort has been driven by a programme of four workshops in North America, Europe and Oceania (Miller and Gill 1997), pilot implementation by a number of extremely enthusiastic groups in Europe and elsewhere, and an active electronic mailing list upon which issues are debated between workshops. A fifth workshop is due to be held in the Finnish capital, Helsinki, in October 1997, where attention will be devoted to some of the many ways in which the scope of Dublin Core has been redefined in order to enable its adoption within large resource discovery projects such as the Nordic Metadata Project (1997) and the Arts & Humanities Data Service (Miller and Greenstein 1997).

The definition of the Dublin Core has evolved somewhat since the first workshop in Ohio in 1995 (Weibel *et al* 1995), but is felt to have stabilised - for now - with fifteen elements, each of which is importantly both optional and repeatable, and potentially refinable by means of optional qualification with 'SCHEME', 'SUBELEMENT' or 'LANGUAGE' information.

The fifteen elements are Title, Creator, Subject, Description, Publisher, Contributors, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights, with the official definition of each element available from the Dublin Core's World Wide Web site (Weibel and Miller 1997). An interpretation of these definitions intended specifically for a (UK) Humanities audience is available from the Arts & Humanities Data Service (Miller 1997).

The three optional qualifiers are available in order to allow clarification both of the way in which an element is being used and of the context from which the element value is drawn.

SCHEME is used to identify a controlled terminology or coding scheme from which any value is drawn, allowing the recorder, for example, to identify that dates are drawn from International Standard 8601 and that 1997-06-05 therefore must refer to June rather than May. This qualifier might also be used to identify various specialist thesauri, allowing users to determine the context from within which a particular term was selected.

SUBELEMENT allows the recorder to specify which particular aspect of each of the fifteen elements is being recorded at any given time. The Dublin Core's Coverage element, for example, deals with both spatial and temporal coverage, and allows the entry of text (place names, archaeological periods, *etc.*) as well as various forms of number. Whilst an entry of 'North Lanark' is perfectly acceptable as a coverage, this value becomes more meaningful when qualified with the SUBELEMENT 'placeName' (telling the searcher that 'North Lanark' is the name of a place) or even 'placeName.authority.unitary' (telling the searcher -- and the computer's search engine -- that 'North Lanark' is the name of an administrative Unitary Authority). Further examples of the manner in which SCHEMEs and SUBELEMENTs can be used are available from the Arts & Humanities Data Service (Miller and Greenstein 1997).

Finally, LANG allows the recorder to specify the language in which the *metadata* is expressed. This differs from the Dublin Core's Language element, which defines the language of the *resource*. A copy of Shakespeare's *Hamlet* in the French national library in Paris, for example, may well still be in English (so the Dublin Core Language element will record 'English'), but the cataloguing information -- the metadata -- will most logically be in French, so each Dublin Core element would therefore be qualified by a LANG qualifier recording the value 'fr', the code for French in ISO 639.

5 Resource Discovery versus Content Detail

Although potentially an extremely powerful resource discovery tool, the Dublin Core is by no means capable of replacing entrenched standards such as MARC, CSDGM, or the plethora of smaller standards relevant to archaeology (see Miller and Wise 1997 for a partial list of these), *and nor was it ever intended to*. The Dublin Core -- and any related initiative which follows, arising from the World Wide Web Consortium's work on resource discovery -- is intended *solely* as a resource discovery tool, and the very simplicity which makes it so powerful in this arena renders it almost useless at the more detailed level served by these other standards.

The almost ridiculously obvious, yet architecturally elegant, solution proposed to solve the potential problem of combining subject detail with discipline-neutral generalisation is the Warwick Framework's notion of a metadata 'package' (Lagoze *et al* 1996). Proposed at the

second Dublin Core workshop, and developed further by Carl Lagoze and others, the Warwick Framework makes it possible for a 'package' of generalised Dublin Core metadata (tailored for resource discovery in an interdisciplinary environment) to be associated with a 'package' of extremely detailed discipline-specific metadata in any suitable format (tailored directly to the needs of the host discipline). The user thus interacts with the broadly intelligible Dublin Core description until he/she has found the resource and decided it is what they want, at which point they may view far more detailed information relating to such arcana as book binding types or aerial photographic camera apertures. It is thus potentially possible for archaeological data to be recorded in a wide variety of formats, described using any number of metadata syntaxes, and still uniformly searched by means of a single Dublin Core-based distributed catalogue.

This is the premise upon which the Archaeology Data Service is constructing its catalogue, a prototype version of which should be available from http://ads.ahds.ac.uk/ahds/ on the World Wide Web by the time this paper is published.

6 Conclusion

Deployment of metadata throughout archaeology is by no means a panacea for all of the profession's failings but -- especially as the potential for reuse of data within archaeology and beyond continues to grow -- documented adoption of resource description and discovery standards cannot help but increase the reuse value of all those kilometres of rotting paper, wrinkling mylar, and demagnetising magnetic media in rarely visited archives around the world.

For far too long, the archaeological community has demonstrated a depressing degree of insularism in insisting upon barely adequate home-grown solutions to problems rather than adopting and adapting more generic solutions from the world outside our trenches, and our documentation procedures are no exception.

A large number of initiatives in disciplines other than our own have much to offer -- and we, too, have much to offer them. An integrated approach to the recording of metadata about their -- and our -- work is an important step towards closer integration in the future. Let's not throw this opportunity away by adopting a home-grown solution to metadata recording, but rather take part in the ongoing international efforts and shape these emerging standards in such a way that they meet *our* needs, as well as the needs of others.

Bibliography

ISO 639 - "Code for the representation of names of languages", Geneva: International Organisation for Standardisation

ISO 8601 -- "Data elements and interchange formats -- Information interchange -- Representation of dates and times", Geneva: International Organisation for Standardisation

Federal Geographic Data Committee, 1997 Content Standards for Digital Geospatial Metadata April 1997 Draft , http://www.mews.org/nsdi/revis497.pdf

Lagoze, C, Lynch C A and Daniel R, 1996 The Warwick Framework: a container architecture for aggregating sets of metadata , http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell%2fTR96-1593?abstract=warwick

Library of Congress, 1997 MARC home page , http://www.loc.gov/marc/

Miller, P, 1996 Metadata for the Masses, Ariadne, 5 , http://www.ukoln.ac.uk/ariadne/issue5/metadata-masses/

Miller, P, forthcoming *Metadata: an aid to interdisciplinary resource discovery*?, Computers and the Humanities

Miller, P, 1997 Unifying Resource Discovery Metadata for the Humanities: an application based upon the Dublin Core, in Miller, P, Greenstein D (eds), *Discovering Online Resources across the Humanities: a Practical Implementation of the Dublin Core,* Bath, United Kingdom Office for Library and Information Networking, 34-55

Miller, P and Gill T, 1997 Down Under with the Dublin Core, Ariadne, 8 , http://www.ariadne.ac.uk/issue8/canberra-metadata/

Miller, P and Wise A, 1997 Resource Discovery Workshops: Final report from the Archaeology Data Service , http://ads.ahds.ac.uk/ahds/project/metadata/workshop1_final_report.html

Miller, P, Greenstein D (eds), *Discovering Online Resources across the Humanities: a Practical Implementation of the Dublin Core,* Bath, United Kingdom Office for Library and Information Networking

Nordic Metadata Project, 1997 The Nordic Metadata Project home page , http://linnea.helsinki.fi/meta/index.html

Weibel, S, Godby J, Miller E and Daniel R, 1995 OCLC/NCSA Metadata Workshop Report , http://www.oclc.org:5047/oclc/research/publications/weibel/metadata/dublin_core_report.html

Weibel, S and Miller E, 1997 Dublin Core Metadata Element Set: Reference Description . http://purl.org/metadata/dublin_core_elements

Weibel, S, Iannella R and Cathro W, 1997 The 4th Dublin Core Metadata Workshop Report, D-Lib Magazine, June 1997 , http://hosted.ukoln.ac.uk/mirrored/lis-journals/dlib/dlib/dlib/june97/metadata/06weibel.html

Contact details

Paul Miller
Collections Manager
Archaeology Data Service
King's Manor
York YO1 2EP
UK
email: collections@ads.ahds.ac.uk
www: http://ads.ahds.ac.uk/ahds/