Excavating the Hard Drive: Archaeological Research, XML, and 3D Graphics

Thomas L. Milbank
Perseus Digital Library
Tufts University
124 Eaton Hall
Medford MA 02155
tmilbank@perseus.tufts.edu

Abstract

Whether running custom or commercial software, computers play an increasingly important role in the recording of textual, graphic, and spatial data on archaeological excavations. However, the data are often compromised by the proprietary formats in which they are stored. Such formats are not archival, and they pose problems for information retrieval. The problems are surmountable within the controlled and limited environment of an active excavation. But outside these confines, the formats become unwieldy and an impediment to large and diverse data sets like those created through federation. The development of federated data sets requires data to conform to a standard. International standards are already advanced for electronic texts that have a history of SGML, and more recently XML, applications. Consequently text passages can be indexed and selectively retrieved from large, tagged corpora. In contrast, the application of XML to standards-based 3D graphics files is immature. Similar tools for indexing and retrieving select graphic elements from large corpora are lacking.

This paper focuses on the X3D XML application of the VRML international 3D graphic standard. It addresses the integration of X3D graphic files into the Perseus XML document management system. And it addresses the creation of a tool to extract and represent graphic elements from multiple files. The tool provides a specific research mechanism for the discovery of embedded graphic data interspersed through a large corpus, for example, allowing an archaeologist to retrieve every cornice from an extensive collection of files without having to open and search each file manually.

1. Archaeological research requires a data format that is permanent, flexible, and easily indexed

Computers are an indispensable tool for modern archaeological research. As archaeologists, we rely on a variety of software to record, analyze, and store the data that we collect. Many of us have tried to access or repurpose old data but, to our displeasure, found them unusable. These experiences demonstrate that proprietary binary formats, however ubiquitous and convenient, are not suited to our discipline's fundamental need for data preservation. Instead, we require formats that are permanent, flexible, and easily indexed. In general, we better serve our needs by storing our data in text-based formats that adhere to international standards [1].

2. The Extensible Markup Language (XML) offers a suitable format for archaeological data

The Extensible Markup Language (XML) is one text-based format that is increasingly popular [2]. It is a subset of the Standard Generalized Markup Language, an international standard for the representation of electronic texts with human-readable markup. In XML, electronic data are stored as strings of text enclosed within descriptive tags that, collectively, define the structure of the data. The combination of plain text and descriptive markup makes XML significant for archival purposes. Using any text editor, an XML-encoded data file can be opened, the file's structure and content discerned, and new programs written to access the data. The explicit nature of XML and the use of standard XML formats provide data with permanency and flexibility. Moreover, XML's text basis makes it easy to index the data in a file; indices in turn make it possible to locate undocumented information in the data. Indexing our data is analogous to excavating

them, in so far as it reveals and records material that was previously hidden.

3. 3D archaeological data encoded in the Extensible 3D (X3D) format have all the benefits of XML

XML is primarily a format for encoding electronic texts, but it is also a format for encoding the data managed by a number of widely different applications. In fact, we can use XML to store 3D archaeological data in a way that meets all of our format requirements. Extensible 3D (X3D) is one of several standard XML formats available for 3D graphics [3]; specifically, it is an XML application of the Virtual Reality Modeling Language (VRML). A 3D graphic encoded in X3D inherits the permanency and flexibility characteristic of a community-based XML format. It likewise inherits XML's ease of indexing. The ability to index graphic data is important since we can expect our interest in the different parts of models to change with shifts in the direction of our research. Without indices, we are limited to hand-keyed metadata that describes the whole of models when locating 3D archaeological data to evidence our work. This limitation is a non-trivial barrier to resource discovery, especially within large data sets that have multiple authorship or other aspects of heterogeneity.

4. When research concludes, migration of X3D-formatted data to an institutional repository is trivial

Standard XML formats provide 3D archaeological data with the qualities we require for our own research. They also prepare the data for eventual integration with large repositories. The data from research conducted by previous generations of archaeologists have found their way into institutional archives; we should anticipate a similar store for the electronic data that we currently produce. Digital archives and libraries are in rapid development at academic institutions across the country, and standard formats are central where there are tools to support the data. 3D graphics encoded in the X3D format will migrate readily from the restricted context of a research project to the open context of an institutional repository. The format's characteristics allow repositories to parse the encoded 3D data, with the result that individually modeled but undocumented geometry can be identified and even extracted. Ultimately, the use of XML to format our 3D archaeological data will benefit the research of archaeologists who come after us.

5. An architectural model of an Egyptian tomb demonstrates the power of the X3D format

The model is a reconstruction of tomb G 2110, a Fourth Dynasty mastaba, and is part of a larger model of the Western Cemetery of Giza. The model captures both published and unpublished archaeological data recorded during the Harvard University–Boston Museum of Fine Arts excavations between 1932 and 1938 [4]; it exemplifies one archaeological application of 3D technology to reexamine old data for new insights. The model is encoded in XML-based X3D, so migration from its project context to an XML-aware repository is trivial. The process neither requires nor causes modification of the data.

The X3D Document Type Definition describes the elements that define the navigation, lighting, view and other components of the mastaba model. The `<Shape>` and `<Group>` elements are among the elements that define the model's geometry, including that of a sculpted door jamb now in the Boston Museum of Fine Arts (Boston 07.1002). The Perseus XML document management system parses the mastaba model, maps the various geometry labels to an "object" abstraction, and generates an index of the normalized data. The index captures all identifying information contained in element `DEF` attributes and includes the byte offset for each component of the model. At run-time, the document display routine references this index, finds the extents of specific geometry, and extracts it. This geometry is then merged with additional tags to create a new X3D document. Present practice passes the new document to an Extensible Stylesheet Language (XSL) transformation utility for styling into VRML before it is displayed in a web page.

With the mastaba's X3D file in the system, an archaeologist seeking information about the sculpture "Boston 07.1002" can retrieve either a 3D model of the relief alone or a 3D model of the relief in its context. These results are notable in light of the mastaba model's hand-keyed metadata, which give no indication that the sculpture is included among the model's sub-objects. The results are possible because of the index created by the document management system, and they highlight the importance of specificity in a model's internal metadata. Standard layer naming conventions provide some general internal metadata, but other means of documenting a model's geometry should also be used to capture particulars.

6. The Perseus Digital Library is developing new research technologies that build upon XML-enabled geometry extraction

The ability to isolate component geometry is a starting point for the integration of XML-based models with automatic linking tools, whereby cross-references will be inserted into 3D data. And it presents an opportunity to explore enhanced 3D shape-matching services, whereby models' constituent shapes might be indexed and searched.

7. Acknowledgments

A grant from the Digital Libraries Initiative Phase 2 (NSF IIS-9817484) provided support for this work.

8. References

[1] International standards relevant to this poster include:
Extensible Markup Language/Standard Generalized Markup Language (ISO 8879)
Extensible 3D (ISO/IEC 19775)
Virtual Reality Modeling Language (ISO/IEC 14772)

[2] Archaeological projects that implement XML include:
The Giza archives project
http://www.mfa.org/giza/
The Hacimusalar project
http://www.choma.org/public/hm/home
The Stoa
http://www.stoa.org/
The XSTAR project
http://www-oi.uchicago.edu/oi/proj/xstar/xstar.html

[3] Additional standard XML formats for 3D graphics include:
CAD-3D
http://www.web3d.org/CAD-3D/
ifcXML
http://www.iai-international.org/iai_international/

[4] Reisner, George A. 1942. *A History of the Giza Necropolis*. Vol. 1. Cambridge: Harvard University Press.