ASOR 2010

Open Context

Prepared by:
Eric Kansa and Sarah Whitcher Kansa

The Alexandria Archive Institute

GUIDELINES FOR WEB-BASED DATA PUBLICATION IN ARCHAEOLOGY

A working document to inform archaeologists about sharing data, from the field to the Web.

Last Revised: Nov. 14, 2010

Table of Contents

Section I – Glossary of some Useful Terms

- **"Machine-Readable" Data:** Most content on the Web is oriented for human consumption via Web browsers. However, software is playing an increasingly important role in aggregating relevant content from different sources on the Web to make discovery, analysis and visualization easier. By publishing "machine-readable" data using formats that are better suited for software to process, you greatly expand options for finding and using your content.

- **Metadata:** More information specialists define metadata as "information about information". Metadata is used for a variety of purposes, sometimes for discovery of information. Library catalogue information including fields for authors, titles, publication data, and subjects represent familiar examples of discovery metadata. Other forms of metadata can be used to describe the meaning of information. For example, a description of the meaning of a field in a database would be a form of metadata.

- **Semantic Web:** The Semantic Web includes a suite of concepts and technologies that add metadata to Web resources. The point of adding these metadata is to more precisely define the content on the Web. For example, is the term "Lot" the name of a person (say a Biblical character), or a type of contextual unit used in archaeological recording? Semantic Web metadata could be used to remove such ambiguity. While in theory the Semantic Web sounds like a good idea, in practice it is often very difficult for people to align themselves to a common standard for meaning. Thus the Semantic Web mainly sees implementation in settings where there's a great deal of consensus (certain areas of the "hard sciences") or where consensus can be enforced by policy (government agencies).

- **REST:** "Representational State Transfer". REST is not a standard but a style for designing information systems. Most of the standards behind the World Wide Web adhere to RESTful design patterns. REST design styles see great interest because of their inherent simplicity, especially for sharing data. For example, in RESTful design patterns, the only thing you need to do to obtain an information resources is to go to its address or location. On the Web, this means following URLs to *GET* information. Scholarly websites designed according to RESTful principles are easier to use, easier to reference, and easier to integrate with other sites on the Web.

- **Creative Commons:** In many respects, copyright law works against scholarly interests, particularly when it comes to publishing reusable data on the Web. By default, copyright law prohibits all reproduction and reuse of content without express permission by the copyright owner. This restriction defeats the purpose of data-sharing[1]. Creative Commons licenses are useful in that they expressly remove default copyright restrictions to provide clear permissions for reuse of content. Creative Commons therefore enhances *legal interoperability* of online data.

[1] Copyright laws distinctions between "facts" (not protected) and "expressions" (protected) make matters even more complicated. It's sometimes hard to tell if your data looks legally factual or expressive.

Section II – Some Simple Web Design Tips

1. **Use Open Standards.** Unless you have no alternative, it is best to use nonproprietary open standards for scholarly applications. Data expressed in open standards can be used on more computing platforms and they are much easier to preserve and archive.

2. **Make Valid Web Resources.** Many websites fail to use standards properly. They are more likely to load and display improperly on some browsers. They are also more likely to not work well in the long-term. The World Wide Web Consortium (W3C) has a free validation service (http://validator.w3.org/) to test your content. Please ask your Web developers to use it!

3. **Keep it Simple**. Simple websites are more likely to last and see support on the widest variety of platforms. Simple HTML or XHTML should be preferred over proprietary formats like Adobe-Flash. If you use more sophisticated dynamic content (using lots of scripting), be sure to design your site for "graceful degradation"), so that it can be usable even if a particular component fails to work properly.

4. **People are not your Only Audience.** Interoperability and reuse are key issues. Computer software is an increasingly important "audience" for your content. This is critical for making your content reusable and easier to migrate to new platforms, thus improving the longevity of your content. From the beginning, you and your web developers should design your site so that it provides data for both people and machines. It's good to be aware of current *open* standards for facilitating interoperability and data aggregation by machines. The more widely used the standard, the more likely other people and machines will be able to use, understand, and preserve your data.

5. **Remember, Hyperlinks are Important.** Content published on the Web should be easy to reference with a hyperlink (URL). It sounds simple, but URLs are one of the most important aspects of interoperability and usability on the Web. They are also important for discovery. If all of your content can be found by following links, Google and other search engines can also find your content. Also if people have an easy time linking to different parts of your site, your PageRank (impact) can improve.

   A key issue for designing a site with longevity is good and thoughtful design of URLs. Good URLs should be short, and should avoid indications of backend scripting languages or other implementation details (in other words, no ".php" or ".asp" should be in a URL). This gives site owners the ability to keep existing URLs even if they change technologies on their servers. A given resource should have one URL, and it shouldn't vary depending on the login-status of the person retrieving the information.

Examples

Some examples of **BAD** URL design[2]:

- British Museum, London. http://www.britishmuseum.org/research/search_the_collection_database/search_object_details.aspx?objectid=117631&partid=1
- Metropolitan Museum of Art, New York. There is a database search form at http://www.metmuseum.org/search/woa_advanced_search.asp but it does not seem to support individually addressable catalog entries via a URL.

Some examples of **GOOD** URL design:

- Arachne. http://arachne.uni-koeln.de/item/bauwerk/2100006
- Portable Antiquities Scheme, UK http://finds.org.uk/database/artefacts/record/id/412455
- Nomisma.org (Ancient Mediterranean coin hordes) http://nomisma.org/id/athens
- Pleiades (Gazetteer of the Greco-Roman world), http://pleiades.stoa.org/places/462503

[2] Examples and discussion by Sebastian Heath, see: http://wiki.digitalclassicist.org/Citation_of_Archaeological_Collections_and_Aggregators and http://wiki.digitalclassicist.org/Citation_of_museum_collections

Section III – Questions to ask of a Digital Resource

1. **How stable looking is the URL?** If the URL has lots of different query parameters in it, or indications about a server's technical features (".aspx" or ."php", which relate to scripting languages not content), they you can be pretty certain that the URL will likely not work in a few years. Well designed URLs are more likely to be more stable.

2. **How would you cite this Resource or Item?** A hyperlink is the most important way resources are "cited" on the Web. If the Web resource indicates some sort of backing by digital libraries, URLs will likely resolve in the future. If something has a permanent identifier associated with it (like a DOI) then it's much more likely to be scholarly and citable with digital library support (like a journal article), meaning that it will be retrievable in the future.

3. **Can you tell who the Author is?** Is the resource attributed to an individual or institution that has some credibility? Credibility is somewhat in the eye of the beholder. One way to determine credibility is to look for links from trusted sites. Links usually indicate some sort of vote of confidence. You can use services like Yahoo's "site explorer" (http://siteexplorer.search.yahoo.com) to check who links to a particular website.

4. **Is there Quality Control?:** Scholars typically trust journals because journals adhere to known and accepted processes for quality control. Such processes can also take place in online contexts. However, scholars often jump to conclusions about open access online scholarship and erroneously believe Web resources must inherently be poor in quality. Just because something is available on the Web (often for free) does not mean it's not peer reviewed or vetted. When looking at a Web resource, check for documentation about quality control and vetting.

5. **Don't Judge a Book by its Cover:** The site may be beautiful, but if it's all built in Flash, it's not going to be sustainable (or able to be archived). The more open standards it uses, the more permanent and reliable it will be.

Section IV – Preparing your Data for Web Publication

This section provides a series of simple steps you can take to make your project data ready for web publication in Open Context (http://opencontext.org) or another data publication system. Providing the following information and quality tests will ensure that your content is archived and understandable by scholars years from now using altogether different technologies than those which you used to create your data set.

Step #1: Clean it Up!

- Translate all coded items
- Standardize your entries as much as you can (for example, do not have "Bos" and "bos"—choose one and stick with it)
- Check spelling
- Make sure each row has a unique identifier (such as "Specimen Number")
- Provide a brief description of what each field header means
  - Example 1: "Basket: Sub-context of a locus in which the specimen was deposited."
  - Example 2: "Bd: A measurement representing the breadth of the distal end, as defined by von den Driesch (1976)"
- Prepare a separate spreadsheet listing each image name and the item (or locus, area, or site) it represents (each image should have its own row)

Step #2: Provide Project Information (Metadata)

In order to facilitate archiving and reuse of your work, it is critical that you provide information about your project data. This "metadata" will be forever attached to your dataset and will ensure that future users of your data will have access to key information about your project and your methodologies. The following list is in decreasing order of importance. While providing information on all criteria is ideal, the first five criteria are imperative to make sense of a dataset.

Critical

- **Project Title**
- **People**
- **Abstract (detailing project aims, methods, and recording practices)**
- **Geographic coverage (site location)**
- **Chronologic coverage (overall time period, in calendar dates or cultural period)**

Useful but not Critical

- Keywords
- Funding agency and/or sponsoring organization
- List of participants and their biographies
- Reference resources used
- Where does the physical collection current reside?
- Citations of related published literature and links to web resources

See example here: http://opencontext.org/projects/B1DAC335-4DC6-4A57-622E-75BF28BA598D

Step #3: Publish a Synthesis and Link it to the Data

In addition to sharing your research data, always try to also produce a narrative interpretive publication that draws on those data. The publication should describe the research questions that guided your work. Ideally, that publication should be made available open access and linked to your published dataset. This ensures that the primary dataset and interpretations based on it will always reference each other, allowing for more informed future uses of the dataset and readings of the synthetic publication.

Additional resources on this topic

Extensive guidelines on documenting datasets can be found in *Digital Archives from Excavation and Fieldwork: Guide to Good Practice Second Edition* (produced by the Archaeology Data Service and the Arts and Humanities Data Service and edited by Julian Richards and Damian Robinson): http://ads.ahds.ac.uk/project/goodguides/excavation/

This online document also provides specific guidelines about *Documenting the Digital Archive*: http://ads.ahds.ac.uk/project/goodguides/excavation/sect42.html

Section V – Structuring a Data Access Plan

In 2010, the National Science Foundation (NSF) announced new data sharing requirements for grantees. Grant-seekers now need to provide as part of their proposals details of their plans to ensure access and long term preservation of their project data. This new requirement has the potential for improving transparency in research. Shared data also opens the door to new research programs that bring together results from multiple projects.

The NSF archaeology program links to Open Context (http://opencontext.org) as a data archiving service to help applicants meet the requirements of the Data Access Plan. Open Context offers researchers guidance on how prepare datasets for presentation and how to budget for data dissemination and archiving (with the California Digital Library). Grant-seekers can use Open Context's online estimation form (see here: http://opencontext.org/about/estimate), which upon submission provides them with a cost estimate and the following information (by email) that they may incorporate into their grant applications:

> *"Digital content generated by this project will be published in Open Context (http://opencontext.org). Publication in Open Context ensures that the data are freely available and openly licensed so they can be reused and combined with research collections from elsewhere on the Web. All project data will be documented with relevant metadata ("information about information"), citation information, and a permanent URL to maximize reach and potential for reuse.*
>
> *All content published in Open Context is archived by the California Digital Library, a leader in the preservation of scientific data. These features enhance the quality, discoverability, and interoperability of project data beyond what can be achieved by simply posting data to a website.*
>
> *The following provides additional detail about other features and design attributes Open Context provides to maximize your project's digital content:*
>
> > *(1) Deep Linking: Every item of the project dataset in Open Context will have its own Web Address (URL). This policy of "one webpage per potsherd" enables very specific referencing and citation of Open Context content. Citation is further encouraged by dynamic generation of suggested citation text, and by support of metadata standards used in Zotero, a popular reference management tool. In addition, deep linking enables Web 2.0 users to tag project content with Delicious.com or similar folksonomy services or, others can apply sophisticated and more formal semantic standards.*
> >
> > *(2) Open Access: Open Context requires no login to access, download, or copy data into another system. Its stated policy to refrain from monitoring individual user*

activities is consistent with the American Library Association's code of professional ethics to protect patron rights to privacy, confidentiality, and academic freedom. The absence of a login barrier also allows Open Context content to be fully indexed by commercial search engines, further enhancing discoverability and impact.

(3) Machine-Readable Data and Services: Open Context data and querying services come in a variety of data formats, services, and protocols (Atom, GeoRSS, KML, ArchaeoML/XML, JSON, CSV, OAI-PMH). These measures help ensure that project content can flow into other applications that may visualize it in new ways or combine it with data from other sources.

(4) Open Licensing: Project data will be released under the Creative Commons (http://creativecommons.org) copyright licenses. These licenses open the door to future research, instruction, and other applications. These standard licenses explicitly grant permissions for reuse of content, provided attribution is given to content producers. To facilitate interoperabity, the licenses are expressed as standard, machine-readable metadata, using the RDFa format. Creative Commons licenses and Open Context's machine-readable data help insure this project's content can be moved to other applications and archives. Because content is not "trapped" in Open Context, Open Context will enable, not inhibit, future uses of these data as new systems emerge."

Section VI – A Few Key Web Standards

Standard	Purpose	Description
HTTP	Web communications	The "Hyper-Text Transfer Protocol". This is the communications protocol used on the World Wide Web. HTTP defines the Web as a subset of the larger Internet, where other protocols are also often used (FTP, email, etc.).
HTML / XHTML	"Human-readable" (with a browser) Web content	HTML is the "hyper text markup language" and "XHTML" is a more formal version of HTML. These are content standards for Web pages. HTML and XHTML define special "tags" to indicate stylistic as well as certain general semantic properties of content on a Web page. For example, the tag "<h1>" describes a heading, meaning that words within this tag should be interpreted as a title.
XML	"Machine-readable" data	XML is the "extensible markup language". XML is a more general standard than HTML or XHTML. HTML and XHTML are standards with pre-defined tags while XML is a standard that allows you to define your own tags. This allows you to custom-tailor and precisely define the meaning of data you are sharing. Typically, a community (an academic discipline, or another group needing to share data) defines a set of XML tags to facilitate sharing of data. The Archaeological Markup Language ("ArchaeoML", defined by David Schloen for the University of Chicago OCHRE Project) represents one such example. GML (the "Geographic Markup Language") is another widely used implementation of XML for sharing Geographic Information Systems (GIS) data.
RDF	"Machine-Readable" data (Semantic Web)	RDF (the "Resource Description Framework") has many similarities to XML. RDF is used to precisely define the meaning and relationships within content so that data can be processed by software. While XML is best used to define "tree"-like data structures (hierarchic relationships), RDF is best used to describe "graph" or "web"-like data structures. Because of this, RDF is the favored standard for publishing machine-readable data for the Semantic Web. Unfortunately, it is also harder for programmers to use.
JSON	"Machine-Readable" data (enriched user interactions)	JSON ("Javascript Object Notation") is a simple, light-weight standard for expressing machine-readable data. It is mainly used to provide data to widgets and applications that power richly interactive features on certain websites. For example, GoogleMaps can read JSON data to display custom map overlays on a website.
Atom	"Machine-readable" data (syndication format)	Atom is a very versatile XML format for syndicating (sharing) content on the Web. It is generally regarded as the best designed version of RSS (Really Simple Syndication) formats. Atom can be used for sharing notifications of updates of content, or even for sharing lists of resources retrieved from a query. Both such applications are very useful for data sharing.
KML	"Machine-readable" data (mapping and geospatial visualization)	KML is a widely used XML format for sharing geospatial data. Although it is mostly associated with GoogleEarth and GoogleMaps, KML is a nonproprietary open standard that is also supported by Open Layers (an open source project), Yahoo Maps, and Microsoft's "Bing Maps".
COinS or unAPI	"Machine-readable" data for bibliographic citations	Zotero (http://zotero.org), a popular free and open-source citation management tool (plugin for the Firefox browser) supports a number of standards for automatically reading bibliographic citation information from a web-page. COinS and unAPI are two examples.