Data Curation Profile – Architectural History / Epigraphy

Profile Author	Christopher Eaker
Author's Institution	University of Tennessee, Knoxville
Contact	Christopher Eaker, Graduate Research Assistant School of Information Sciences University of Tennessee, Knoxville ceaker@utk.edu
Researcher(s) Interviewed	W
Researcher's Institution	University of Tennessee, Knoxville
Date of Creation	May 2, 2012
Date of Last Update	
Version of the Tool	1.0
Version of the Content	1.0
Discipline / Sub-Discipline	History of Architecture / Epigraphy
Sources of Information	• An initial interview conducted on April 26, 2012. • A second interview conducted on April 30, 2012. • A worksheet completed by the researcher and interviewer • Pogany, A., Sands, A., Collazo, S., & Lee, Y. (2012). *Data Management Final Project: Visualizing Statues in the Late Antique Roman Forum*. University of California, Los Angeles, Los Angeles, CA. • Epidoc website (http://epidoc.sourceforge.net/) • Project website (http://inscriptions.etc.ucla.edu/) • A sample of the profiled data.
Notes	N/A
URL	
Licensing	Creative Commons Attribution-NonCommercial-ShareAlike 3.0

Section 1 - Brief summary of data curation needs

The researcher has provided public access to this data as a service to others within and outside of his field. By providing this access, he hopes to generate discussion with others in his field and provide educational opportunities to students in the fields of archaeology, the classics, and history.

The TEI/XML guidelines employed on this project called Epidoc provides for a useful way to contribute this work to the larger scholarly inscription community, though standards within this community are still not solidified.

http://dx.doi.org/10.5703/1288284314995

Section 2 - Overview of the research

2.1 - Research area focus

The research project is entitled “Visualizing Statues in the Late Antique Roman Forum.” This project focuses on inscriptions from statue bases from the 4th and 5th centuries CE in a specific neighborhood of ancient Rome called the Forum. Some of the inscriptions are extant and some have been lost. Some were recorded during the Renaissance by inscription hobbyists who left manuscripts or published books. There are 95 discrete texts from these inscriptions. Unfortunately, none of the statues survived.

This project encoded the inscriptions using a specific mode of transcribing inscriptions called Epidoc. The data set is a set of TEI/XML encoded inscriptions. Additionally, the researchers attempted to suggest the physical locations of these statue bases based on a variety of information by assigning geographic coordinates to them. The project’s website is http://inscriptions.etc.ucla.edu/.

2.2 - Intended audiences

The people who would be most interested in this data are archaeologists, epigraphers (people who study inscriptions), and classicists (people who study ancient Greece and Rome). The researcher also believes the data are useful for students. The general public may also find this interesting.

While the researcher has not tracked it, he has gone to conferences and met people who have found it very useful, namely archaeologists and people interested in classical architecture in ancient Rome. The work was originally intended for a scholarly audience, but other groups are conceivable.

2.3 - Funding sources

The project that produced this data was funded primarily by the National Endowment for the Humanities, but also by the University of California, Los Angeles, and the University of Tennessee, Knoxville.

NEH did not require the researchers to create a data management plan, though they did, in fact, create one. Nor did the NEH require that they share or preserve the data, but they both share and preserve it.

Section 3 - Data kinds and stages

3.1 - Data narrative

Initially, the text and photos of the various statue inscriptions were collected from published sources. If a photo of the statue base was not available, the photograph was taken in the field. Next, the statues were assigned GPS coordinates using a combination of GPS devices and Google Earth. Next, the texts were encoded into TEI/XML using the Oxygen XML Editor. The number of complete data files is 95.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Collecting Raw Inscription Data	Texts of inscriptions	95 / <1 MB each	TXT, XML	Collected from published sources
Encoding Inscriptions in TEI/XML	Encoded TEI/XML files of inscriptions	95 / <1 MB each	TEI/XML	Encoded in Oxygen XML Editor; full XML files are not provided on the project website
Final Published Data	Selected parts of encoded TEI/XML files of inscriptions	95 / <1 MB each	HTML, MySQL	Published on the project website; Only the inscription part of XML files are provided on the project website
Ancillary Data				
Digital Photographs	Photographs of statue bases	95 / ~1 MB each	JPG, TIF	Added to provide photograph of the actual inscription on the project website
GPS Coordinates	GPS coordinates of inscriptions	95 / <1 MB each	KML	Added for displaying the statue locations on a Google Earth map within the project website

Note: The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray.

3.3. - Target data for sharing

The data was shared only with immediate project collaborators in the initial data collection phase and in the encoding inscriptions phase. Once the data were finalized, the data was shared publically on the project website. However, the full TEI/XML files are not publically available on the project website due to a database limitation. Only the inscription text portion of the XML file is publically available because it was copied and pasted into the database. However, the researcher is glad to share these files with anyone who requests them.

3.4 - Value of the data

This data has value to multiple groups. One benefit is access to the text by those interested in inscriptions from this historical period. Another way these data add value is that they demonstrate how these texts had a living function in ancient Rome. Furthermore, the researchers made assertions that these texts have meaning based on their locations. They believe that it would be interesting for other scholars to respond to the arguments they put forward about the social implications of the statues' locations.

3.5 - Contextual narrative

The inscriptions were encoded in a specialized form of TEI/XML called Epidoc. This standard was developed by researchers at the University of North Carolina, Chapel Hill. Epidoc encodes the semantic elements of the inscription text, not the physical appearance. Epidoc follows the Leiden Conventions, which determined how epigraphs should be displayed in modern texts. For example, hard brackets around a piece of text (i.e. [abc]) indicates that the text was missing from the original and has been provided by the epigrapher. Another example is parenthesis around a piece of text (i.e. (abc)) indicates that there was an abbreviation in the original text that has been spelled out by the epigrapher.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)
In response to the question, “Who is the owner of the data?” the researcher replied, “I have no idea, and I really wouldn’t claim it to be me.” There is little to this data that he feels a sense of ownership over since he derived the texts from publically available publications. The researchers offer the website to the community as an interesting portal to this information.

4.2 - Stakeholders
The main stakeholders in the production of this data were the project funders -- NEH, UCLA, and UT, and other collaborators, such as Diane Favro from UCLA and four graduate students who helped the researcher compile and encode the 95 inscription texts.

4.3 - Terms of use (conditions for access and (re)use)
The researcher does not place any conditions on the access and use of this data.

4.4 - Attribution
The researcher would appreciate attribution for the value he added through the GPS coordinates and TEI/XML encoding. However, he understands that there are standard methods of citing inscriptions, and if someone cited his work using the formally recognized methods, then no plagiarism would have taken place. Therefore, citation of his work by others is not a high priority for him.

Section 5 - Organization and description of data (incl. metadata)

5.1 - Overview of data organization and description (metadata)
The researcher assigned a descriptive title for each file. The title describes the object which the inscription describes. For example, if the inscription described Emperor Constantine on horseback, the title was "Equestrian Statue of Constantine." In addition to this titling convention, the researcher assigned metadata within each XML file, such as the title of the inscription, the GPS coordinates, the Corpus Inscriptionum Latinarum number, and revision history.

5.2 - Formal standards used
The data is encoded in a formal TEI/XML language for encoding inscriptions called Epidoc. Epidoc is standardized and its website contains documentation and guidelines on its use (http://www.stoa.org/epidoc/gl/5/toc.html).

5.3 - Locally developed standards
Not discussed by the researcher

5.4 - Crosswalks
Not discussed by the researcher

5.5 - Documentation of data organization/description
Not discussed by the researcher

Section 6 - Ingest / Transfer

The data is currently accessible on a UCLA server; therefore, the researcher does not see the need to transfer it to a repository, institutional or otherwise, because this would essentially be duplicating the data. The researcher would be willing to submit the data to a disciplinary specific repository geared towards inscriptions if there were a suitable one available. However, the

various groups of people around the world who work on inscriptions are not coordinated in their methods and do not use a standard method of encoding inscriptions.

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

Most of the data is currently available on the project website, so it is apparent that the researcher is willing to share the data publically. However, the full XML files are not accessible. The researcher would be willing to share this data with anyone who asks.

7.2 - Embargo

Not applicable

7.3 - Access control

The researcher does not have any need to control or restrict access to this data.

7.4 Secondary (Mirror) site

The researcher would find it convenient to access the data from a secondary site if the repository is offline and would place a medium priority on it.

Section 8 - Discovery

The researcher places a high priority on the ability for researchers in his discipline and outside his discipline to easily find the data set. Though he is not sure how interested the general public will be in the data, he believes it is a high priority for them to be able to find the data set because he believes it is important to generate public support.

He also places a high priority on the ability of people to easily discover this data using internet searches and believes this is probably the primary way people will discover it.

Section 9 - Tools

Tools used to generate the data are the following:

- Hardware: GPS devices, digital camera, PC and Mac computers
- Software: Oxygen XML Editor, Google Earth, Google Spreadsheets, MySQL

Tools used to access, use, visualize, and interpret the data are the following:

- Hardware: PC or Mac computer
- Software: Mozilla or Chrome web browser with Google Earth plugin (Internet Explorer does not work), Text editor or Oxygen XML Editor to view XML files.

The researcher places a high priority on being able to visualize the data. This is the reason the data is placed on a project website along with visualization tools.

Section 10 – Linking / Interoperability

The researcher places a high priority on linking his data with publications. He places a medium priority on supporting the use of web service APIs. He places a high priority on being able to

merge his data with other data sets, but believes this is logistically difficult for reasons outlined in Section 6.

Section 11 - Measuring Impact

11.1 - Usage statistics & other identified metrics

The researcher places a low priority on the ability to see usage statistics. He wants to have as many users as possible and may be interested in monitoring usage, but he thinks quality of use is more important than quantity of use.

11.2 - Gathering information about users

The researcher places a low priority on obtaining information on the people who make use of the data. He would, however, be interested in knowing how other people are using the data.

Section 12 – Data Management

12.1 - Security / Back-ups

The data set is hosted on a server at UCLA and is backed up consistently. However, the researcher's personal backup practices are less frequent and consistent. He backs up twice a year. He does not utilize any other security measures.

12.2 - Secondary storage sites

The researcher would place a high priority on having the data hosted on a secondary site at a different geological location, but believes it is not always feasible to do so.

12.3 - Version control

The researcher does not believe version control is important for this data.

Section 13 - Preservation

The researcher sees the benefit in preserving all parts of the data for this project, including the XML files, the digital photographs, and the GPS coordinates. They are all interrelated and interlinked; therefore, they should be preserved as a group.

13.1 - Duration of preservation

The researcher believes the data should be preserved indefinitely since the content of the data is already close to two millennia old. The researcher said, "Since people have been preserving it this long, we might as well try to keep it just as long."

13.2 - Data provenance

Not discussed by the researcher

13.3 - Data audits

The researcher places a low priority on the ability to conduct data audits.

13.4 - Format migration

The researcher places a high priority on the ability to migrate the data. This is the reason he used XML for encoding the inscriptions. He believes XML is very easy to update and transfer.

Section 14 – Personnel

14.1 - Primary data contact (data author or designate)
Withheld from the public version of the Data Curation Profile.

14.2 - Data steward (ex. library / archive personnel)
Not discussed by the researcher

14.3 - Campus IT contact
Not discussed by the researcher

14.4 - Other contacts
Not discussed by the researcher