6. Data Management Plan

6.1 Nature of the Data Produced

This project will generate archaeological data expressed in a variety of open formats. *Open Context* offers a host of RESTful services for access to machine-readable data. These services are fully documented with live examples at: http://opencontext.org/about/services. All query results from *Open Context* are expressed as:

- *JSON*: The Javascript Object Notation format makes it easier for third parties to develop applications using *Open Context*'s querying services.
- *Atom Syndication Format*: Atom is a widely used standard ideal for sharing lists of resources. *Open Context* expresses all query results as paged Atom feeds, using the feed-paging and archiving link relations. Each Atom entry expresses links to XHTML+RDFa, XML (ArchaeoML, see below), JSON, CSV, and RDF representations of all of *Open Context*'s content. These Atom feeds make it easy for third parties to discover and retrieve multiple machine-readable representations of *Open Context* content.

The California Digital Library's Merritt digital repository crawls *Open Context*'s Atom feeds. *Open Context*'s Atom feeds provide an updated list of resources for Merritt to discover and accession. The Atom entries allow Merritt to discover and archive *all* machine-readable representations of *Open Context* data (and associated media files, including images). Thus, Merritt archives *Open Context* data in several different open and nonproprietary formats, better ensuring the long-term usability of *Open Context*'s content.

6.2 Metadata and Standards

Open Context uses a variety of standards to promote interoperability with domain-specific cultural heritage systems and more general digital library systems. *Open Context* represents all archaeological data using the Archaeological Markup Language (ArchaeoML), documented at the University of Chicago OCHRE project (Schloen 2001). In addition, *Open Context* provides XML-RDF data conforming to the CIDOC-CRM. CIDOC-compliant data representations are based on the pioneering work of Claros, a JISC-funded museums interoperability initiative documented at (http://www.clarosnet.org/wiki/index.php?title=CIDOC_CRM_Objects).

In addition to domain specific standards, *Open Context* expresses metadata using a variety of more widely used standards common to digital library systems. These standards include Dublin Core and MODS. MODS metadata can be retrieved using the unAPI protocol and is primarily useful for generating bibliographic metadata for the Zotero citation management tool. Dublin Core metadata is expressed as XML in all ArchaeoML documents using Dublin Core's XML namespace.

Open Context and the California Digital Library are in active development finalizing integration of Merritt archival services, with a March 2012 completion date (see the *Open Context* blog, Heritage Bytes for updates: http://ux.opencontext.org/blog). *Open Context* also expresses Dublin Core metadata in Atom, which facilitates integration of Merritt's archival services. In crawling *Open Context*'s Atom feeds, Merritt gains publication and update information, as well as other Dublin Core metadata needed to accession content into the Merritt repository. Upon accession into the Merritt repository, Merritt mints persistent identifiers. The persistent identifiers used by *Open Context* include DOIs and ARKs. Because DOIs are more expensive but have more "prestige value," they are used to identify entire datasets (at the project or data table level). ARKs, on the other hand are free, and are assigned individually to every record in *Open Context*, enabling highly granular and specific citation of *Open Context* content. DOIs and ARKs comply with DataCite requirements for persistent identification of resources, enabling *Open Context* to use the DataCite standard. With DataCite, citations to *Open Context* data can be tracked using digital library infrastructure and services, eventually enabling future determinations of citation impact factors.

6.3 Linked Data Entities

In addition to the various metadata and representation standards discussed above, *Open Context* references entities published by external data sources using Linked Open Data methods. Currently, these vocabularies include GeoNames (for geographic entities), Freebase (for units of measurement), Pleiades (for ancient places), and the Encyclopedia of Life (for biological taxa). Pleiades already has archival support of the NYU library system, and the Encyclopedia of Life has long term support of the Smithsonian Institution, Harvard University, and others. To provide additional archival support, we will accession records of entities referenced by datasets in this project to the California Digital Library. All of the external data sources referenced by this project provide Creative Commons licensed data, which facilitates these archival efforts.

6.4 Versioning

Version control for datasets will begin in pre-publication editing, cleanup and documentation stages. Google Refine, a major component of the *Data Refine System* is used to coordinate edits, tracks versions and enables roll-backs of changes. The entire history of data edits is stored, and can be retrieved and archived as a compressed JSON file. This revision history can be archived along with the final published version of a dataset in Merritt. However, data contributors may have privacy concerns and may not want to archive early draft stages of their data. Thus, we will only require archiving of finalized datasets and leave archiving of draft versions of data to the discretion of contributors.

The Merritt digital repository stores a representation of all versions of content accessioned from *Open Context*. *Open Context*'s Atom feeds express timestamps for all new content and updated or modified content. The Merritt repository's Atom crawler uses this information to either accession new content or update previously archived content, creating a new version with the current updated information. Thus, Merritt ensures the continued accessibility and preservation of earlier versions of data published with *Open Context*.

6.5 Licensing

To ensure wide interoperability and free reuse of data, all data published in *Open Context* for this project will be licensed with the Creative Commons Attribution License.