Data Publication
Guidelines for Contributors

Overview

Open Context is a Web-based data publication system that offers free access to editorially verified datasets, many of which are linked to print publications. Open Context also offers an optional peer review process to further validate datasets. The system uses a simple and generalized data model that can accommodate most archaeological datasets and museum collections. Because archaeological databases vary widely in organization and in terminologies, we have developed a specialized application (called “Penelope”) to import data into Open Context for Web publication.

While we have worked hard to make importing data into Open Context simple, imports still require some understanding of the ArchaeoML data model behind Open Context. Therefore, Open Context editors are the main users of Penelope and they import data on behalf of contributors and work closely with data authors to ensure that the data are published in a logical and appealing way. In the future, we hope to further automate the import process, eventually making it easier for users to “self-publish” their data. However, as Open Context’s content is editorially verified, some editorial involvement will always be required.

While Penelope is very flexible and can accommodate most datasets, some preparation is required before importing data. Most of this preparation involves documenting a dataset so that it can be understood by Open Context’s editors and future users of the content, including other researchers and students. This document describes how to prepare your data for smooth publication in Open Context.

Data Preparation

1. **Good Database Design:** Good database design from a project’s start makes eventual data publication easier. Normalization (removal of redundant information) helps to maintain data quality. Maintaining consistency is also important. For example, numeric data fields should contain only numeric data. If additional notation or explanation is required for some numeric information, these should go in other fields. Data validation (error checking) practices throughout data collection will speed data publication and help make the published data more valuable and easier to use by others.
2. **Clean-up and Edits:** Publishing data in Open Context is a form of publication, but one that differs from journal articles or books. Because datasets are often fairly “raw,” one should not expect perfect spelling, grammar, or compositional excellence in daily logs, database comment fields, etc. Spelling problems in these fields will probably have little impact on the overall usability of contributed data. However, some errors have greater impact. For instance, nominal values (terms used over and over again), such as the terms used to describe artifacts in a small finds database (“lamp,” “coin,” “spindle-whorl”), should be consistent (in terms of plurals, terminology, and spelling) to aid search and understanding. Identifiers for objects or contexts (such as “catalog #,” “locus #”), especially those that have associated descriptive information should also be free of errors.
3. **Decoding:** To speed up data entry, many people use coding systems as a convenient way to record data. However, these coding systems may be unintelligible without explanation. To facilitate understanding of a dataset, our preferred practice is to request that data contributors replace code with intelligible text before import.
4. **Description and Explanation:** Each field of every dataset must have some narrative description to

aid interpretation, even if only a sentence or two. Sometimes certain values in these fields should also be described, especially if data contributors employ terminology that is not widely used by their colleagues.

5. **Structural Relations:** Archaeologists often manage their data in relational databases with complex structures. These structures need explanation so that editors can perform the proper queries to “extract” data. Specifically, we will need to know the primary and secondary (foreign) “keys” in each table.
6. **Locations and Objects:** Open Context creates a separate web page (retrieved at a URL) for every location and object, person, and media file it publishes. It is important to let Open Context editors know which fields represent identifiers for different locations and objects (archaeological sites, archaeological contexts, survey tracks, artifacts, ecofacts). Ideally, some descriptive information should be made available for each identified location and object, including excavation areas or trenches, even if these descriptions are only in narrative form.
7. **Images and Media:** Images and other media comprise an important component of archaeological documentation. Each individual media file must be clearly and unambiguously linked to one or more specific records in the dataset (such as records of excavation contexts, people, excavation log records, artifact records, etc). The data contributor should prepare a separate table listing each image file name, an image description (if desired) and the number / identifier of the object or place the image describes.
8. **Abstract and Background:** Each project should have a narrative abstract or background description. This should provide introductory information describing the project goals, key findings, as well as methods and recording systems. For large projects, contributors can also provide additional supplemental background descriptions of specialist analyses. These materials may be submitted in Microsoft Word (or similar) format.
9. **People and Attribution:** For citation purposes, every record in Open Context must be attributed to one or more specific person(s). In some cases, certain database fields have records of different people who made observations and analyses. Ideally, the people identified in these fields should be identified by full name (not initials) and these names should be spelled properly. In other cases, entire data tables or datasets are created by a single person (such as a specialist). For each data table, please provide a name and institutional affiliation for the person(s) primarily responsible for authorship.
10. **General Good Practice:** The UK-based Archaeology Data Service offers Guides to Good Practice (http://ads.ahds.ac.uk/project/goodguides/g2gp.html). These are currently being revised and expanded in collaboration with Digital Antiquity. Open Context editors highlight these documents because they are invaluable guides to improving data quality, longevity, and usability.

Data Formats and Structures

Data for import should be in Microsoft Excel tables. The first row (“row 1”) of the table should contain data field names (columns). The other rows should have the data records in the table, with each data record listed in a separate row. If you do not have Excel or cannot produce Excel spreadsheets from your database, Penelope can also handle Filemaker, Access, and Open Office, as well as comma separated value files. Please note however, that you first must extract image and other media from a database (if stored in “binary fields”) and store them as individual files.

The project abstract/background should be in Microsoft Word (or a similar format). In addition to the above, you may also provide as much supporting or related documentation as you like, such as PDFs of related publications, extended bibliographies in Word, and links to related web resources (such as descriptive project web sites, profiles of project participants on their institutional web sites or links to self-archived publications related to the dataset).

Location Information and Site Security

Open Context requires at least one geographic reference for each project. This geographic information should be most pertinent location information useful for interpretation. This is usually the location of sites. Because Open Context makes all data freely and openly available, data contributors must consider site security issues associated with revealing location data. If location data represents a threat to site security, these data should not be revealed with great precision. Instead, sensitive location information should be randomized and only provided publicly at reduced precision. Users should be informed of this manipulation, and contact information needs to be provided for qualified researchers to obtain precise location data. Please contact Open Context's editors if you are concerned about site location.

Copyright and Licensing

Open Context publishes open access, editorially controlled datasets to support future research and instructional opportunities. Thus, data contributors must make their content legally usable by others. To ensure legal reuse, we require that all content be released to the public domain, or that contributors use Creative Commons (creativecommons.org) copyright licenses on their content. We strongly recommend users select the Creative Commons "Attribution" license. The Attribution license is easiest to understand and helps makes contributed data widely useful as possible. While we allow licenses that restrict commercial uses, we recommend against such restrictions. Please be aware that such restrictions are inherently ambiguous and would inhibit important uses, such as inclusion of content in textbooks or even journals distributed through sales.

In the US, copyright applies to expressive works, not compilations of factual information. Therefore, Creative Commons copyright licenses are not appropriate for some datasets, especially those with limited "expressive" content. Datasets that are less expressive and have less "authorial voice" tend toward a more "scientific" and factual nature (i.e. those that mainly include physical measurements and adhere to widely used conventions in nomenclature and recording). These datasets should use the Creative Commons-Zero (public domain) dedication.

We encourage contributors to choose a single license to apply to the entire dataset; however, we can also assign different license choices to individual items.

Please note that copyright and licensing issues are largely independent of scholarly citation and attribution. *Professional standards dictate that all users properly cite data contributors even for public domain content, especially for scholarly uses*. This professional norm of conduct works independently of the copyright status of content.

Publishing Fees

To support open access publication and archiving, Open Context has developed a pricing structure based on a contributor-pays model. Publication fees vary between $250 and $6000 depending on the complexity and size of the contributed database and related content. For example, a single spreadsheet of faunal data, with no related images would cost on the low end of this spectrum. In contrast, a complex project with several databases, specialist analsyses, and thousands of media files, would be on the high-end of this scale. Open Context developers can provide additional fee based services for implementations based on Open Context's Web-services (API) or other customizations. To assist in budgeting, interested contributors should contact the editorial team (see below) to establish a fee for their specific project.

Grant Seekers (NSF and Other Granting Programs):
We have developed a form to help plan your data sharing requirements. This form will help you budget appropriately for data sharing and it will generate text you can use for the "Data Access Plan" section of an NSF proposal. Other granting agencies may wish to see similar plans. Once you successfully complete the form, you will receive an email with a budget estimate and language to add to your Data Access Plan. This language will include a description of the access, interoperability, and archiving issues. This form is available here: http://opencontext.org/about/estimate

Peer Review

Open Context content must pass professional editorial scrutiny before it can be published on the web. Open Context only accepts content from professional / accredited researchers, government officials, and museum staff. Contributors must be able to demonstrate adherence to appropriate legal, ethical, and professional standards of conduct and methodological rigor. In addition to these editorial controls, contributors may also request peer-review of their contributions. If peer review is desired, please provide names and contact information for possible reviewers, as well as a brief description of the expertise needed to make informed judgments of your content. Content that passes peer-review will be clearly marked as such on Open Context.

Questions?

Please address questions to the Editor of Open Context, Sarah Whitcher Kansa (skansa@alexandriaarchive.org).

[This document was last updated: March 30, 2011]