Designing Metadata for Long-Term Data Preservation: DataONE Case Study

Betsy Gunia
Northwestern University Library
1970 Campus Dr.
Evanston, IL 60208
betsy-allen@northwestern.edu

Robert J. Sandusky
University of Illinois at Chicago
801 S. Morgan St.
Chicago, IL 60607
sandusky@uic.edu

ABSTRACT
DataONE (Data Observation Network for Earth) aims to ensure the preservation of and access to multi-scale, multi-discipline, and multi-national earth observation data to enable advances in science and science education. DataONE is being designed and built to manage scientific data across a range of disciplines, including atmospheric, ecological, hydrological, oceanographic and other earth sciences, all of which are managing data and creating metadata at varying levels of maturity and complexity, utilizing dozens of metadata standards. This poster describes how PREMIS (Preservation Metadata: Implementation Strategies) was utilized to specify the requirements for preservation metadata for DataONE, one aspect of DataONE's technology architecture.

Keywords
Data curation, digital preservation, metadata, preservation metadata, PREMIS.

INTRODUCTION
DataONE's mission is to "enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it" (http://dataone.org/about). The project has numerous goals, one of which is to build a globally-distributed network of member nodes and a smaller number of coordinating nodes. Member nodes can participate in DataONE by registering and implementing a published application programming interface (API)[1]. Member nodes can be based on any hardware and software platform and typically store a mix of locally generated data and copies of data replicated from other member nodes. Coordinating nodes are responsible for supporting DataONE-wide services including data discovery and access, ensuring data integrity, and coordinating preservation actions. Well-designed preservation metadata is necessary for DataONE to preserve scientific data for the long term. In the project's first year the development efforts have focused on defining and implementing the elements of the API needed to establish the first set of member nodes and the first three coordinating nodes. This examination of preservation metadata requirements has proceeded in parallel.

The DataONE technical architecture is designed to allow scientists to create and deposit data described by any of the dozens of metadata standards and file formats already in common use in particular scientific disciplines. Each of these descriptive metadata standards may include varying amounts of structural, technical, preservation, and administrative metadata. DataONE has defined the first iteration of DataONE system metadata: a lingua franca, added as an additional metadata layer, to support the discovery, use, reuse, and management of DataONE's digital content. DataONE's first iteration of preservation metadata is also included in the system metadata.

As scientific data is added to DataONE, data services extract selected descriptive elements from the metadata provided by the scientist to support search, faceted discovery, and other common DataONE services. DataONE also assigns additional system metadata values to support the DataONE distributed architecture, authentication, authorization, data and metadata replication and synchronization, use statistics, preservation services, etc.

PRESERVATION METADATA ANALYSIS
The DataONE systems metadata (sysmeta) must include information that supports the long-term preservation of these objects, but the first year iteration is incomplete. We examined PREMIS in order to identify the gaps that needed to be filled to create a comprehensive preservation metadata scheme. We compared the sysmeta specification, created by the Core Cyberinfrastructure Team (CCIT), with the PREMIS Data Dictionary and then identified gaps between the two. PREMIS suggests many additional semantic units not covered in the year one sysmeta specification, like information on the operating environment, file storage location, preservation policy and digital signature

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

[1] The details of the DataONE architecture are stored at http://mule1.dataone.org/ArchitectureDocs/.

information. Also, PREMIS and the systmeta specification deal with derived and new versions of digital objects differently. For example, if object 1 is a previous version of object 2, PREMIS links these objects through the relationship semantic unit, while the sysmeta specification uses the field "Obsoletedby" to hold the unique identifier of object 2. Finally, except for recording the date and time when an object was created and verified against its checksum, the sysmeta specification does not have metadata fields for all of the suggested activities that PREMIS records (such as migration and compression).

We found two reports on applying PREMIS to scientific data. The National Snow and Ice Data Center (NSIDC) originally created their own schema for preservation metadata, but after learning about PREMIS, tried to reconcile the two schemata and apply PREMIS to their datasets (Duerr, Weaver, & Parsons, 2010). The authors drew three lessons from their research. First, scientific datasets are comprised of smaller units called "granules," i.e., the smallest units upon which metadata is managed. Consequently, metadata about datasets and granules are stored separately in NSIDC's data systems. PREMIS, on the other hand, accommodates all levels of units, resulting in extremely large tables and calling scalability into question. Also, if a dataset is transformed to a new format and the scientific integrity is unaffected, NSIDC classifies the dataset as a new version and applies the prior metadata to the new version. PREMIS, instead, assumes that metadata is immutable and relevant to one and only one object. Finally, NSIDC did not implement the preservation rights section of the Data Dictionary because the first version of PREMIS did not define preservation responsibilities pertinent to scientific data centers. The authors concluded that, at least at the representation level, PREMIS would be useful for earth science data.

The National Geospatial Digital Archive (NGDA) also reported on their analysis of the utility of PREMIS (Hoebelheinrich, N. & Banning, J., 2008). They compared and contrasted how FGDC CSGDM, PREMIS, and GER applied to four different geospatial data types. They concluded that although PREMIS lacks descriptive metadata, it has two strengths: 1) application to both abstract and physical components of a resource through the relationship unit, and 2) ability to record actions taken during the object's lifecycle in the repository. The authors suggested that a combination of FGDC CSGDM and PREMIS is sufficient for the long-term preservation of geospatial datasets.

CONCLUSION

The differences between the sysmeta specification and PREMIS, and conclusions from other examinations of PREMIS-based implementations for science datasets informed our recommendations to the CCIT concerning additional metadata fields. Among the most important recommendations are: (1) include fields on the software and hardware environments needed to render digital objects, (2) expand fields that describe both structural and derivative relationships between objects, and (3) record additional preservation actions. The analysis and literature review of previous PREMIS-based implementations have been conducted and we are in the process of documenting the gaps in a format useful for the CCIT.

Defining the requirements for preservation metadata is only part of the process of creating an effective and sustainable system for data preservation. Metadata containers, such as METS (Metadata Encoding and Transmission Standard), also must inform a final operational design. Standards for aggregations, or packages, of Web resources such as the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) (Lagoze et al, 2008) and the BagIt File Packaging Format (Boyko et al, 2009) must also be taken into account as possible methods for recording the relationships between preserved objects in DataONE published papers and reports.

ACKNOWLEDGMENTS

This work is supported in part by Data Observation Network for Earth (DataONE), NSF award #0830944 under a Cooperative Agreement.

REFERENCES

Boyko, A., Kunze, J., Littman, J., Madden, L., Vargas, B. (2009). The BagIt File Packaging Format (V0.96). Retrieved April 2, 2010, from http://www.ietf.org/internet-drafts/draft-kunze-bagit-04.txt.

Duerr, R., Weaver, R., & Parsons, M. A. (2010). A new approach to preservation metadata for scientific data: A real world example. In L. Di & H.K. Ramapriyan (Eds.), *Standard-based Data and Information Ssystems for Earth Observation* (pp. 113-125). Berlin: Springer-Verlag.

Hoebelheinrich, N. & Banning, J. (2008). An Investigation into Metadata for Long-Lived Geospatial data Formats. Retrieved on July 12, 2010 from http://www.digitalpreservation.gov/partners/ngda/ngda.html.

Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., Warner, S. (2008). Open Archives Initiative Object Reuse and Exchange: ORE User Guide - Primer. Retrieved April 2, 2010, from http://www.openarchives.org/ore/1.0/primer.

PREMIS Editorial Committee. (2008). Data Dictionary for Preservation Metadata: PREMIS version 2.0. S.l. Retrieved April 2, 2010, from http://www.loc.gov/standards/premis/v2/premis-2-0.pdf.