Library Philosophy and Practice (e-journal)　　　　　Libraries at University of Nebraska-Lincoln

2014

PREMIS-Lite, a Preservation Metadata Generator.

Todd P. Swanson
San Jose State University, tddswnsn@gmail.com

PREMIS-Lite, a Preservation Metadata Generator.

Todd Paul Swanson

San Jose State University

Abstract

Digital preservation is without question an important, if not essential, facet of our now digital society. Implementing a digital preservation strategy is crucial to ensuring the longevity and sustained usability of digital content. The importance of digital preservation is not limited to large-scale institutions and organizations, but also extends to private individuals and small-scale businesses.

PREMIS is currently the international preservation metadata standard. The objective of PREMIS was to identify and create a wide ranging and implementable set of "core" preservation metadata elements while following the recommendations of the Open Archival Information System (OAIS) reference model. However, the depth and complexity of that coverage can be confusing and discourage its potential use.

Using the PREMIS standard as a model this project creates a "PREMIS-Lite" xml preservation metadata generator. This PREMIS-Lite XML generator and its supporting PREMIS-Lite semantic unit definition, user guide, and crosswalk combine to allow for the simple capture and creation of preservation metadata at the object level.

Keywords: PREMIS, PREMIS-Lite, Preservation, Metadata, Digital, Generator, XML.

Digital preservation is without question an important, if not essential, facet of our now digital society. Whether one is a digital content creator, manager, or perhaps both, understanding and implementing a digital preservation strategy is crucial to ensuring the longevity and sustained usability of digital content. The importance of digital preservation is not limited to large-scale institutions and organizations, but also extends to private individuals and small-scale businesses as well. Finding ways to adapt standards designed to fit large-scale institutions for smaller organizations is an important task in helping the larger audience of digital content creators, managers, and society as a whole. In order to analyze and adapt such standards it is vital to understand the current state, and history, of preservation metadata.

In June of 2002 an international working group sponsored by the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG) was created to address the growing importance of establishing, creating, and implementing a digital preservation metadata standard. This OCLC/RLG working group used the already established work in preservation metadata developed by the library communities CEDARS (CURL Exemplars in Digital ARchives), NEDLIB (Networked European Deposit Library), and the NLA (National Library of Australia) as a springboard to create the PREMIS (PREservation Metadata: Implementation Strategies) preservation metadata standard.

PREMIS is currently the international preservation metadata standard and is maintained by the Library of Congress Office of Network Development and the Office of MARC Standards. The objective of PREMIS was to identify and create a wide ranging and implementable set of

"

"core" preservation metadata elements while following the recommendations of the Open Archival Information System (OAIS) reference model. The major outcomes from the PREMIS Working Group included, in 2004, a survey *Implementing Preservation Repositories For Digital Materials: Current Practice And Emerging Trends In The Cultural Heritage Community*, the *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*, and its supporting XML schema, which were released in May 2005. Following its release, the PREMIS Data Dictionary won the 2005 Digital Preservation Award and the 2006 Society of American Archivists Preservation Publication Award.

The ongoing work of maintaining, supporting, and coordinating future modifications and revisions to PREMIS is supported by the PREMIS Maintenance Activity, hosted by the Library of Congress. The work of the Maintenance Activity includes a website aggregating and linking official and unofficial PREMIS information, a discussion list, and a wiki for PREMIS implementers, as well as an Editorial Committee. The Maintenance Activity also attempts to promote awareness of PREMIS, sponsors tutorials in using PREMIS, and commissions' studies and publications related to PREMIS

To better understand PREMIS it is first important to understand the building blocks used to create it. The CEDARS Project and resulting *Cedars Preservation Metadata Elements* was one of the key pieces used to help inform and guide the creation of PREMIS. The project ran from 1998 through 2002 and was under the overall direction of the Consortium of University Research Libraries (CURL) and the universities of Cambridge, Leeds and Oxford. The project

goal was to promote awareness of the importance of digital preservation, produce strategic frameworks for digital collection management policies, and promote methods appropriate for long-term preservation.

The CEDARS project specifications are covered in *Metadata for Digital Preservation: The CEDARS Project Outline Specification,* which describes the metadata elements that the CEDARS Project identified as being useful to ensure that digital library resources can be archived and used in the future. However, despite their work of creating an element set the project did not specify or indicate any particular method of implementation. In March 2002, Michael Day produced the *Cedars guide to preservation metadata,* part of a series of guides published by the Project. Day also published event reports of the Cedars conference *Preservation 2000* and (with Maggie Jones) the final *Cedars Workshop***,** held in February 2002. According to the *Cedars Guide to: Preservation Metadata*:

> *At the time the Cedars project proposal was being put together, there was an awareness of the perceived importance of preservation metadata, but there was no existing 'standard' that could be adopted by the project for use in its demonstrator services. The project bid proposed, therefore, that Cedars would produce a metadata specification.*

T

The next building block used to help inform the creation of PREMIS was NEDLIB

(2000). The NEDLIB element set defines a "core" set of essential preservation metadata, with an

emphasis on overcoming the problem of technological obsolescence. The report *Metadata for*

Long Term Preservation (2000), covering NEDLIB 2000, declares its objective as attempting to

define "core minimum metadata that are mandatory for preservation management purposes, in

order to handle large amounts of data items in a changing technological environment." This

"core" includes 8 metadata elements and 38 sub-elements. The report focuses strictly on

preservation related metadata and deems other types of metadata categories such as descriptive,

administrative and, rights out of the project's scope. Additionally, the report doesn't cover every

type of digital document with specificity, rather it targets the most generic information about

digital objects in effort to not limit the metadata sets potential use.

Similar to the CEDARS Project and NEDLIB work from the NLA also helped shape

PREMIS. According the NLA staff paper published in February of 2000 *Preservation Metadata*

for Digital Collections the digital preservation work of the NLA was founded to "make

Australia's cultural heritage available to future generations, as well as to today's scholars

and researchers". The NLA element set was designed to support the preservation of both

digitized and born-digital objects. The element set is comprised of three levels of descriptive

granularity; collection, object, and sub-object. This set is also designed to be, what the staff

paper refers to as "implementation-neutral" meaning that no assumptions were made about

S

specific preservation strategies to be used by any NLA element set adopting repository.
Additionally and appropriately The NLA states:

> *Different types of digital materials, and different archiving*
>
> *systems, will need different metadata support. There may be types*
>
> *of material and processes that are not adequately accommodated*
>
> *by our proposal despite our intentions, and we would welcome*
>
> *feedback.*

The predecessors of PREMIS at the larger institutional project level were not the only significant pieces of work stemming from the library community addressing digital preservation. An influential piece of writing related to digital content, digital libraries, interoperability, and preservation metadata comes from Howard Besser's 2002 work *The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries.* This paper discusses the move from isolated digital collections to interoperable digital libraries. In this work Besser discusses the functions (such as infrastructure, robust metadata, and preservation components) that can be deployed to move a library's digital content from isolated digital collections to interoperable digital library. Besser states:

> *We need to widely employ descriptive metadata for consistent*
>
> *description, discovery metadata for finding works, administrative*
>
> *metadata for viewing and maintaining works, structural metadata*

f

or navigation through an individual work, identification metadata

to determine that one has accessed the proper version of a work,

and terms and conditions metadata for compliance with use

constraints.

A technology watch report *Preservation Metadata* published in 2005 by OCLC, the Oxford University Library Services, and Digital Preservation Coalition address the "current state" of preservation metadata and digital preservation. One of the highlights of the report is its coverage on developing a preservation metadata scheme where it aptly states that a preservation metadata schema should aim to be: comprehensive, oriented toward implementation, and interoperable. Furthermore, the report states "Perhaps more than any other form of metadata, preservation metadata requires planners to "get it right" the first time." The report also covers the challenges of creating distinctions between descriptive, administrative, and structural metadata while then trying to find room for "preservation metadata". Again from the report:

A metadata schema is intended to serve some purpose – in a sense,

it supports some "verb" that the schema is intended to help

accomplish. In the case of preservation metadata, the verb is

"preserve", so when we draw a distinction between metadata and

preservation metadata, it is this verb which helps us draw a

boundary around what is in and what is out of scope. Preservation

m

etadata is descriptive, structural, and administrative metadata that

supports the long-term preservation of digital materials.

After the first release of PREMIS in 2005 much of the writing in digital preservation

naturally began to focus on, and feature, PREMIS. PREMIS introductions, reviews, as well as

implementation and usage guides, began to be created and filter out by those engaged with

library community. Of those writings, first and foremost is the information provided by the

PREMIS Maintenance Activity website which includes the PREMIS Data Dictionary, the

PREMIS XML scheme, as well as a variety of other tools and resources related to using and

implementing PREMIS.

PREMIS is based on the idea of implementation, and *Preservation Metadata for*

Institutional Repositories: applying PREMIS attempts to map the five entity types identified in

the PREMIS Data Dictionary -- intellectual entities, objects, events, agents and rights -- to

potential metadata sources identified in an institutional repository preservation model. A few

other PREMIS guides include: a PowerPoint presentation created by Rebecca Guenther for the

American Library Association *An Introduction to Preservation Metadata and the*

PREMIS Data Dictionary, *Understanding PREMIS* by Priscilla Caplan of the Library of

Congress, and *Preservation Metadata for Institutional Repositories: applying PREMIS*. Another

useful resource for understanding preservation metadata and PREMIS includes a tutorial for

d

digital preservation created by Cornell University *Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Problems.*

Published in October of 2006 by Yale University *Using PREMIS to support preservation of digital assets at Yale* examines Yale University's implementation of the PREMIS preservation metadata model. In March 2006, the Yale University Library adopted the PREMIS model for preservation metadata for its digital assets. A task force was then created to investigate, develop, and propose a modified metadata element set and usage guidelines. *Using PREMIS to Support Preservation of Digital Assets at Yale* provides a chart of elements that PREMIS deems mandatory and what Yale deems mandatory for "base profile" material and "full profile" material.

In 2006 the Australian Partnership for Sustainable Repositories (APSR) published the *PREMIS Requirement Statement Project Report.* The current partners of APSR are the Australian National University, National Library of Australia, University of Queensland, University of Sydney, University of Melbourne, University of Technology Sydney and the Australian Partnership for Advanced Computing. The report specifies requirements for the collection metadata needed for the APSR based off of the PREMIS metadata standard. The report cites recommendations for preservation metadata elements including mandatory elements; recommendations on tools for automatic metadata extraction; a recommended list of supported formats and a draft METS profile for exchanging preservation metadata and functional specifications and use cases for preservation events and event logging.

Rebecca Guenther's *Battle of the Buzzwords: Flexibility vs. Interoperability When Implementing PREMIS in METS* examines how PREMIS and METS (Metadata Encoding Transmission Standard) operate together. More Specifically, Guenther targets library communities who plan on using METS to wrap PREMIS and introduces ambiguities in using PREMIS with METS, which she states, "need to be clarified in a set of guidelines."

Published in D-Lib magazine in 2008 by Brian Lavoie *PREMIS With a Fresh Coat of Paint: Highlights from the Revision of the PREMIS Data Dictionary for Preservation Metadata* describes the PREMIS Data Dictionary 2.0 (2008) revision process, its outcomes, and a summary of the major changes in the new version of the Data Dictionary. The paper also touches on a few other topics from the Maintenance Activity, and briefly mentions some areas for future work.

Exploration of adoption of preservation metadata in cultural heritage institutions: Case of PREMIS published by the American Society for Information Science and Technology is one of the more recent writings on PREMIS. The paper states "PREMIS has been extremely influential on providing a "core" set of metadata elements that support the digital preservation process. However, there is no evidence, in the form of previous research, as to what factors explain and predict the level of adoption of PREMIS". The paper goes on to attempt to identify factors that affect the adoption of PREMIS in cultural heritage institutions. A study conducted in support of the paper employed a web-based survey to collect data from 123 participants in 20 countries as well as a semi-structured, follow-up telephone interview, and reviews those results

in the paper. The results of the study aimed to provide recommendations for researchers and stakeholders engaged in digital preservation and metadata standards development efforts.

The understanding, development, and implementation of preservation metadata standards as a whole, and PREMIS specifically, have been thoughtfully considered and refined as the use and understanding of digital content has continued to developed and evolve. The library as a whole, and its extended communities with a vested interest in digital preservation, has made great efforts to design and implement digital preservation strategies that best fit the needs of their digital content to ensure extended use and longevity of such assets. Examining and developing a simple and easy to use version of PREMIS that can be adopted by smaller institutions or private individuals would be of great value to the society digital preservation intends to serve.

PREMIS was designed to be implemented and managed by digital preservation repositories not digital content creators. However, not all parties interested in digital preservation have the resources to devote to the requirements of maintaining a "proper" digital preservation repository such as those PREMIS implementers are expected to have. To address this divide the goal of this project aims to create an approachable method for smaller institutions, individuals, and digital content creators to capture and maintain preservation metadata for all types of digital content. Using the PREMIS standard as a model this project's intent is to create a "PREMIS-Lite" xml preservation metadata generator.

Most of PREMIS is devoted to describing digital objects. However, the depth and complexity of that coverage can be confusing and discourage its potential use. The PREMIS working group produced five XML schemas to encode the semantic units of the Data Dictionary. The schemas can be used independently, in pieces, or together in one package. The five PREMIS schemas are: Object - a schema for the object entity, Agent - a schema for the agent entity, Event - a schema for the event entity, and Rights - a schema for the rights entity. Because objects are what are actually stored and managed in a preservation repository this project will focus solely on the "Object" schema as its coverage pertains specifically to, as the name suggests, the "Object".

To make the use of PREMIS more approachable and easier to use this project will select essential metadata elements from the PREMIS Object Entity to be used in the PREMIS-Lite metadata generator. Those key elements are: fixity information such as a checksum (message digest) and the algorithm used to derive it, the size of the object, the format of the object, which can be specified directly or by linking to a format registry, the original name of the object, and where and on what medium it is stored. It is through these key pieces of preservation metadata, provided in an easy to use metadata generator, that will allow for a digital object's continued long-term viability and longevity.

To utilize these specific metadata element selections the bulk of this project is dedicated to creating an HTML web ready form that generates user input into a simplified version of the XML PREMIS Object schema. This XML generated output can then either function as a stand-alone digital preservation resource, or be implemented with PREMIS or other preservation

metadata wrappers at another time. Additionally, a list describing the PREMIS semantic

units used in the PREMIS-Lite metadata generator will also be provided to the user and function

as a user guide and crosswalk of the PREMIS-Lite element names to PREMIS.

Preservation metadata can be a powerful tool set used to help ensure the viability and

longevity of digital objects. The PREMIS preservation metadata schema is a particularly

effective tool within that set but it can be difficult for the average user to understand and

implement. To make this digital preservation tool easier to use a stripped down version of the

PREMIS schema was created called "PREMIS-Lite". This PREMIS-Lite XML generator and its

supporting PREMIS-Lite semantic unit definition, user guide, and crosswalk combine to allow

for the simple capture and creation of preservation metadata at the object level. Using this

generator allows the everyday user to capture preservation metadata and help ensure digital

content can be used, accessed, and preserved for years to come.

Product

The product from the above mentioned PREMIS-Lite metadata project consists of a

primary HTML web page "SwansonT_metadataProject.html", which functions as a XML

preservation metadata generator. This Page was written using HTML and Javascript languages to

create a form that enables multiple fields of user input, once processed, to be wrapped with

PREMIS Object schema XML tags and made viewable is a separate in page results box.

To support the use of the metadata generator a PREMIS-Lite to PREMIS crosswalk, semantic

unit definition list, and user guide named "PREMIS_lite.html" is made available through the

metadata generator's primary page. This supplemental HTML page allows the user to easily navigate back and forth from the metadata generator main page to the semantic units definition page.

Similarly, the primary HTML preservation metadata generator page also includes links to the National Archives hosted technical registry "PRONOM", as well as the "Online MD5" website which allows users to generate their digital content's fixity and checksum information. Links to the full version of the PREMIS Data Dictionary, as well as the complete XML PREMIS Object schema are also made available to the user through the generator's main page. Lastly, a PDF "metadata_project_example.pdf" has been included with the submission of the project to provide visual examples of the tagged XML preservation metadata the PREMIS-Lite metadata generator generated. Also included in this PDF are a few examples of how that information may be stored along with the digital content it describes on a user's local drive.

To use the PREMIS-Lite Preservation metadata generator download the files: "SwansonT_metadataProject.html" and "PREMIS_lite.html" to your desktop. "SwansonT_metadataProject.html" is the generator and can be opened and used with a variety of web browsers. "PREMIS_lite.html" is the supplemental page consisting of cross-walk/semantic unit definitions/rationale and can be accessed through the link embedded in the generator. Both files can be opened and viewed separately, but were designed to talk to each other and function together.

References

Alemneh, D., Hastings, S. (2011, February 3). *Exploration of adoption of preservation metadata in cultural heritage institutions: Case of PREMIS*. American Society for Information Science and Technology. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/meet.14504701187

Besser, H. (2002). *The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries*. First Monday*, 7*(6). Retrieved from doi:10.5210/fm.v7i6.958

Caplan, P. (2009, February 1). *Understanding PREMIS*. Library of Congress Network Development and MARC Standards Office. Retrieved from http://www.loc.gov/standards/premis/understanding-premis.pdf

Clifton, G., Langley, S., & Lee, B. (2006, July). *PREMIS Requirement Statement Project Report.* National Library of Australia. Retrieved from https://digitalcollections.anu.edu.au/bitstream/1885/46447/4/presta.pdf

Cornell University Library (2003-2006). *Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Problems*. Retrieved from http://www.dpworkshop.org

Day, M. (1998). *Metadata for Preservation*. CEDARS Project Document AIW01. Retrieved from http://www.ukoln.ac.uk/metadata/cedars/AIW01.html

Day, M. (2001 January). *Metadata: Preservation 2000*. Ariadne: Web Magazine for Information Professionals issue 26. Retrieved from http://www.ariadne.ac.uk/issue26/metadata/

Day, M. (2002, May 5). *Cedars Guide to Preservation Metadata*. UKOLN, University of Bath. Retrieved from http://www.ukoln.ac.uk/metadata/cedars/guidance/metadata.html

Day, M., Jones, M. (2002, April 22). *Cedars Final Workshop summary*. Retrieved from http://www.webarchive.org.uk/wayback/archive/20050410120000/http://www.leeds.ac.uk/cedars/pubconf/umist/finalWorkshopRep.html

Dpconline.org. (2005). *The 2005 Digital Preservation Award*. Digital Preservation Coalition. Retrieved from http://www.dpconline.org/advocacy/awards/2005-award

Guenther, R. (2008, July/August) *Battle of the Buzzwords: Flexibility vs. Interoperability When Implementing PREMIS with METS*. D-Lib Magazine, July/August 2008. Retrieved from http://www.dlib.org/dlib/july08/guenther/07guenther.html

Guenther, R. (2008). *Battle of the Buzzwords: Flexibility vs. Interoperability When Implementing PREMIS in METS.* D-Lib Magazine vol 14 7/8 July/August 2008. Retrieved from http://www.dlib.org/dlib/july08/guenther/07guenther.html

Guenther, R. (2009, January 24). *An Introduction to Preservation Metadata and the PREMIS Data Dictionary*. American Library Association Midwinter Meeting 2009, Intellectual Access to Preservation Metadata Interest Group. Retrieved from http://www.loc.gov/standards/premis/premis-ala-mw09.ppt

Hitchcock, S., Brody, T., Hey, J., & and Carr, L. (2007, January 25). *Preservation Metadata for Institutional Repositories: applying PREMIS.* Retrieved from http://preserv.eprints.org/papers/presmeta/presmeta-paper.html
http://www.dpworkshop.org/dpm-eng/resources.html

Jenkins, C., Russell, K. (1998). *CEDARS Project Details*. Retrieved from http://www.ukoln.ac.uk/services/elib/projects/cedars/

Lavoie, B. (2008). *PREMIS With a Fresh Coat of Paint: Highlights from the Revision of the PREMIS Data Dictionary for Preservation Metadata*. D-Lib Magazine, 14 (5/6). Retrieved from http://www.dlib.org/dlib/may08/lavoie/05lavoie.html

Lavoie, B., Gartner, R. (2005, September). *Preservation Metadata.* Joint report of OCLC, Oxford Library Services, and the Digital Preservation Coalition. Published as DPC Technology Watch Report No. 05-01: September. Retrieved from http://www.dpconline.org/docs/reports/dpctw05-01.pdf

Library of Congress. (2013). *Schemas for PREMIS. PREMIS: Preservation Metadata Maintenance Activity (Library of Congress)*. Retrieved from http://www.loc.gov/standards/premis/schemas.html

Library of Congress. (2013). *Tools for preservation metadata implementation - PREMIS: Preservation Metadata Maintenance Activity (Library of Congress)*. Retrieved from http://www.loc.gov/standards/premis/tools.html

National Library of Australia. (1999). *Preservation Metadata for Digital Collections: Exposure Draft*. Retrieved from http://www.nla.gov.au/preserve/pmeta.html

Networked European Deposit Library. (2000). *Metadata for Long Term Preservation*. Retrieved from http://www.kb.nl/coop/nedlib/results/preservationmetadata.pdf

OCLC and RLG working group. (2005, May). *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group.* OCLC Online Computer Library Center, Inc. Retrieved from http://www.oclc.org/content/dam/research/activities/pmwg/premis-final.pdf?urlm=161394

OCLC/RLG PREMIS Working Group. (2004). *Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community.* Report by the joint OCLC/RLG Working Group Preservation Metadata: Implementation Strategies (PREMIS). Dublin, O.: OCLC Online Computer Library Center, Inc. Retrieved from http://www.oclc.org/research/projects/pmwg/surveyreport.pdf

OCLC/RLG Working Group on Preservation Metadata. (2001 January 31). *Preservation Metadata for Digital Objects: A Review of the State of the Art.* Retrieved from http://www.oclc.org/content/dam/research/activities/pmwg/presmeta_wp.pdf?urlm=161396

Russell, K., Sergeant, S., Stone, A., Weinberger, E., & Day, M. (2000 March). *Metadata for Digital Preservation: The Cedars Project Outline Specification*. http://www.egov.vic.gov.au/pdfs/MD-STR~5.pdf

Stone, A., Day, M. (1999). *Cedars Preservation Metadata Elements:* CEDARS Project

document AIW02. Retrieved from

http://www.ukoln.ac.uk/metadata/cedars/papers/aiw02/

Web.archive.org. (2013). *CURL - Consortium of Research Libraries in the British Isles*.

Retrieved from http://web.archive.org/web/20041215094948/http://www.curl.ac.uk/

Www2.archivists.org. (2013). *Preservation Publication Award.* Society of American Archivists.

Retrieved from http://www2.archivists.org/governance/handbook/section12-preservation

Yale University Integrated Access Council [2nd] Preservation Metadata Task Force. (2006,

October). *Using PREMIS to support preservation of digital assets at Yale.* Retrieved from

http://www.library.yale.edu/cataloging/metadata/pmtf/YalePREMIS.pdf