The Seybold Report

Vol 5, No. 06 • June 22, 2005

No Pain, No Metadata
BY VICTORIA MCCARGAR

The PREMIS report on implementable metadata in a variety of settings shows how difficult it is to preserve digital assets. It's not for the faint of heart, so keep your pain relievers handy. A friend who's an expert in taxonomy and vocabulary development mentioned recently that she was considering printing T-shirts that say, "No pain, no metadata" - a terse commentary on how much hard work really goes into creating this vital "data about data."

In fact, a little pain might be good for you. If you're involved in digital asset management (DAM), digital archives or a long-term data retention policy, it's worth a couple of aspirin to sit down with the final report of the Preservation Metadata Implementation Strategies (PREMIS) committee and start to wrap your brain around what kind of metadata is required to keep data for a long time, and where you're going to get it.

The 237-page "Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group" (www.oclc.org/research/projects/pmwg/premis-final.pdf) is the result of almost two years of effort by an international team of digital preservation experts from libraries, archives and the private sector. Hosted by the Online Computer Library Center of Dublin, Ohio (home of Dublin Core), and the Research Libraries Group of Mountain View, Calif., PREMIS was charged with coming up with an "implementable" set of metadata for digital preservation that could be used in virtually any setting. Given the diversity of repositories attempting to keep data "forever," it was a formidable task. But with the interest in digital preservation growing all the time, there was also a sense of urgency.

Rocket Science. Concerns about preserving knowledge in digital form have been mounting steadily since the early 1990s, when the problems were first raised about maintaining digital files amid rapid technological obsolescence. If the ensuing discussion started to sound like rocket science, it's because that's what it was. The antecedents of PREMIS and other data preservation projects are largely found in a document called the Open Archival Information System Reference Model (OAIS), developed by none other than NASA's Space Data Systems group. (NASA is notorious in preservation circles for having lost so much mission data to technology obsolescence.)

OAIS, which became an ISO standard in 2002, in turn spawned a working group (also headquartered in Dublin) charged with looking for the types of information about information - metadata - that would be required to sustain digital files over time. It was more theoretical than practical, however, and lacked workable strategies that could be turned into real systems. PREMIS itself was created to look for them, as well as seek areas of consensus among

Also In This Issue:

different kinds of repositories since what works for a library might not necessarily satisfy a museum's requirements, and neither, of course, would work for a profit-driven newspaper or magazine.

Regardless of the type of repository, sustaining digital information across decades and locations requires a deep knowledge of the environment, components and relationships that go into a digital object, because all of them have a bearing on whether or not that object will survive. PREMIS systematically examines the software, hardware and intellectual environments that produce digital objects, as well as relationships and various dependencies associated with the object. An XML document is useless without its Document Type Definition (DTD), for example. In addition, PREMIS looks at events in the lifecycle of a digital object, such as the deposit of an object into an archive and which entities - human or machine - are authorized to initiate the events. Finally, the schema considers all the information necessary to enable rights management.

Looking for common ground. Given that every repository has unique characteristics, metadata development has to address a huge universe of possibilities. Part of PREMIS' mission - one that was pain-inducing for participants - was to boil the possibilities down into a core set that would cut across all the possible uses in any institution needing to preserve data. The metadata thus avoids "implementation-specific" criteria. It is left to individual institutions to apply PREMIS and expand on it according to their needs. But one of the virtues of adhering closely to the data set, obviously, is the ability to swap preserved data seamlessly with other systems - certainly a criterion in long-range DAM.

The core elements of preservation are spelled out in detail in the PREMIS data dictionary, downloadable separately from the final report at www.oclc.org/research/projects/pmwg/premis-final.pdf. The dictionary is divided into metadata for four "entities" based on PREMIS's data model: the object itself, events in its life cycle, the agents involved in those events and the rights assigned to the object. Each of those entities has an extensive subset of metadata to describe it. The terms assigned to each entity are referred to as "semantic units," which function as a fairly intuitive set of labels. Imagine an XML schema with tags containing the names of the semantic units, such as or . Thus, a small chunk of the metadata that looks at the computing environment is rendered like this:

Software
swName
swVersion
swType
swOtherInformation
swDependency
Hardware
hwName
hwType
hwOtherInformation

The concept behind describing the Object Entity is to include enough information about it to allow future user access, even if the software or hardware that it originally ran on is long since

company's release version of the Window's OS, code-named "Longhorn."

obsolete. These future users would, presumably, be able to use the sum total of Object Entity information to employ one or more retrieval or rendering strategies, including migrating the obsolete data through a proven path, writing an emulator to mimic the original computing environment, or even to fire up the appropriate piece of obsolete hardware if it has been carefully stored. It's likely to be more difficult than just matching up old hardware and software, but PREMIS is ready with encryption algorithm information if one has been used - and if the digital object has one or more relationships to other objects (for example, a Photoshop JPEG to an article written in Word, and that document to a PDF), those can be accounted for in the PREMIS schema.

Events in the life of the object are described as well. Typically, it is important to validate that an object has been successfully ingested into the archives, so there is a set of semantic units to register that outcome. Other preservation activities, such as migrating a file from one version to the next or reformatting a Word file into a longer-lived PDF, can also be noted in the set of Event Entity descriptors. In PREMIS, the people or systems that make those events happen are also duly recorded in the Agents Entity fields.

Such information can also be important to backtracking through problems. Interestingly, we have encountered a situation at the Los Angeles Times where the PREMIS metadata, had it been captured, would be useful for untangling a mystery involving an unknown number of archival JPEGs that were apparently corrupted in an ordinary storage migration several years ago and only recently discovered. Unfortunately, we lack documentation on version and platform changes, or what actions were taken and by whom, so we can only guess at what occurred and have no way of isolating just the affected files for possible repair. Not knowing precisely when the errors occurred, moreover, we have to investigate a two-year time period for possible problems, consisting of about 25% of the database contents, or some 200,000 files.

The PREMIS Rights Entity metadata allows assignment of copyright, links to and terms of a copyright agreement (such as a freelance contract) and specific permissions granted for individual objects. This, again, is to ensure that future users can readily discover their ability to reuse, repurpose or share the preserved digital object. A vague, missing or nonexistent rights statement is as big a threat to the future usability of a digital file as obsolete technology.

Simple but complex. The PREMIS data dictionary comprises fewer than 130 semantic units, which is surprisingly few considered the task they're given: the description of an entire technological environment. The data dictionary is straightforward and fairly easy to understand. It includes suggested rules for applying the units, some instances of how to apply the metadata, and for some semantic units, short taxonomies for assigning real values to a data field.

For the uninitiated, the most useful and least painful part of the PREMIS report is probably the set of examples provided by the team. It offers several cases of the PREMIS metadata applied to real objects from the repositories of team participants (www.oclc.org/research/projects/pmwg/premis-examples.pdf). The

examples range from a simple Word document to a Web site with multiple hyperlinked pages. (I provided an example of a newspaper page produced in Quark with an embedded Macromedia Freehand EPS and a link to the native Freehand original.) The Entity types, hierarchical categories and a raft of data values are presented in an easy-to-navigate spreadsheet format, and each example includes a detailed description of the nature of the object and the computing environment where it resides.

Back to this idea of "no pain, no metadata." The art of applying the PREMIS schema is not for the faint of heart, but, realistically, I think the best way to understand the principles behind PREMIS is to arm oneself with the data dictionary, the set of examples and a file on a local hard disk, and try to describe an object yourself. I asked my UCLA library and information science graduate students to do just that a few days before the PREMIS report was released to the public. They muttered a little about headaches, but except for some legitimate complaints about the level of jargon in the PREMIS documentation, they seemed to grasp the big-picture concepts and populate quite a few of the basic fields in the spreadsheet. But they also found the process laborious, tedious and sometimes confusing.

Their reaction isn't surprising. It has been clear to PREMIS participants from the outset that using human beings to populate the metadata values would be virtually impossible; no one has the time or resources to do it manually, so it wouldn't get done. The possibilities of parsing relevant data automatically or applying predetermined constants programmatically were accounted for throughout the semantic units, and the difference between "mandatory" and "optional" fields often turned on whether a human was required to intervene or not.

Developing automated systems is high on the list of the next steps for the PREMIS project as repositories begin to undertake real implementations. Besides tackling the issue of automation, the near-term future of the PREMIS effort will focus on:

- testing and refining the data dictionary in real repositories;
- development of XML schemas, which is already under way (see www.loc.gov/standards/premis/ schemas.html); and
- ongoing maintenance of the data dictionary and XML schemas (www.loc.gov/standards/premis/)

For the time being, PREMIS will remain the domain of research libraries, which will continue to be the first implementers of digital preservation repositories, especially under the OAIS model. Watch for incorporation of PREMIS into existing standards, such as Technical Metadata for Still Images, exploiting the semantic nesting power of XML. But the rest of us with important legacy material (news archives, for example) are well advised to pay attention to developments in the field even if it's just a little bit painful. It'll hurt a lot more when that legacy material starts going away.

About the Author. Victoria McCargar is involved in newsroom and library technology support and strategic planning at the Los Angeles Times, where she is a senior editor. A frequent lecturer, she served on the PREMIS committee and is currently investigating

digital preservation issues and best practices for the newspaper industry under the auspices of InterPARES, a digital preservation research consortium based at the University of British Columbia. She is also an adjunct professor of preservation at UCLA. She can be reached at mccargar@mac.com.