Summer 1-1-2002

"Steering by Standards": Videoconference Series Report

Ellen McGrath
University at Buffalo School of Law, emcgrath@buffalo.edu

"Steering by Standards" : Videoconference Series Report

by Ellen McGrath
University at Buffalo

"Librarians face the challenge of charting organizational directions that incorporate emerging knowledge-sharing standards. These complex new standards will significantly impact librarians and other information professionals, their institutions, budgets, staffs, systems, and workflows. Gaining practical insights into these global initiatives will help library leaders decide more quickly and effectively how to respond locally. To support this effort, the OCLC Institute is presenting a series of three satellite videoconferences."

This quote describes the overall theme to the "Steering by Standards" videoconference series presented by the OCLC Institute. Details are available at its Web site: *http://www.oclc.org/institute/events/sbs.htm*

I. The first in this series of videoconferences was presented on March 26, 2002 and was entitled "A New Harvest: Revealing Hidden Resources with the Open Archives Metadata Harvesting Protocol." The host was Lorcan Dempsey (Vice President and Director, OCLC Research); the key speaker was Herbert Van de Sompel (Director, e-Strategy & Programmes, the British Library and soon to be at the Los Alamos National Laboratory); and the expert practitioners were Stephen Pinfield (Academic Services Librarian, University of Nottingham, United Kingdom) and Joann Kaczmarek (Project Coordinator, Illinois Open Archives Initiative Metadata Protocol Harvesting Project and Visiting Assistant Professor, Library Administration, University of Illinois at Urbana-Champaign).

The conference opened with an exchange taking place between Jay Jordan (President, OCLC) and Clifford Lynch (Executive Director, Coalition for Networked Information). This set the stage for the presentations to follow by emphasizing the importance of metadata and "the hidden web." Lorcan Dempsey expanded on these concepts in his introduction. Libraries must provide service in new ways in order to tap this wealth of information on the web and showcasing these resources can have a profound effect upon scholarly communication.

Herbert Van de Sompel, one of the founders of the OAI in 1999, then provided a history of the OAI. It grew out of the need to provide a common interface to electronic pre-print resources, so that a user would not have to know of the existence of each archive as well as its search protocol. It then sparked interest outside of the pre-print communities and was released as an experimental protocol in 2001. A second, hopefully more stable, version of the OAI Metadata Harvesting Protocol (MHP) is due out in June 2002.

Van de Sompel explained that there are two sides to the equation: the repository (or data provider) and the harvester (or service provider), each represented by a piece of software. The goal is to provide federated services so that a variety of repositories can be combined into one sort of "super" metadata collection with one search protocol. To illustrate this, Van de Sompel showed a slide with the examples of an e-print archive, a library OPAC, a full-text database, an abstracting and indexing database, and an images archive. The OAIMHP would effectively combine them into one huge database so as to facilitate searching by the end user. The OAI is not at this end point just yet, but it is moving steadily in that direction.

The core concepts of the protocol as outlined by Van de Sompel are:

- Low barrier of interoperability (it is simple)
- HTTP-based
- Replies can be validated against XML
- Replies are self-contained, including the original request
- Shared metadata format and parallel community-specific metadata formats

(Continued on page 7)

"Steering by Standards"

(Continued from page 7)

There was a question and answer session at the end of the formal presentations. Some were technical and related to how the upcoming version of the protocol would differ from the current version and about how high volumes of records are harvested. None of the projects described focus upon the integration of metadata with the library OPAC, but Pinfield said Nottingham was looking into making its OPAC OAI-compliant.

During his presentation, Pinfield had mentioned that catalogers view the Dublin Core as a decline in quality. But he said that he feels the tradeoff is worth it and can be substituted for through keyword access. There was a follow-up question about this and whether it would perhaps lead to the elimination of catalogers. The speakers seemed to agree that at this point some simple metadata access was better than nothing, but that over time a better balance in metadata standards must be achieved.

I had read a little about the Open Archives Initiative (OAI) in the library literature. But I had not really focused in on it closely enough until this videoconference, which I felt was an excellent description of it. It struck me that librarians have traditionally been the "harvesters" in a way. But the overwhelming surge of information available on the Web makes it impossible for librarians to keep pace. This automated solution recognizes that fact and strives to fill the gap.

II. The second videoconference in the series was presented on April 19, 2002 and was entitled "The OAIS Imperative: Enduring Record or Digital Dust?" The host was Meg Bellinger (Vice President, OCLC Digital and Preservation Resources); the key speaker was Donald M. Sawyer (Lead, Science Office of Standards and Technology, NASA); and the expert practitioners were Bruce Ambacher (Electronic & Special Media Records Divisions, US National Archives and Records Administration) and MacKenzie Smith (Assistant Head of Technology, MIT Libraries). This conference opened with Jay Jordan (President, OCLC) conversing with Kathryn Sullivan, one of the astronauts who worked on the Hubbell Telescope.

Don Sawyer, referred to as "the father of the OAIS (Open Archival Information System) standard," had entitled his presentation "Framework for Digital Archiving: OAIS Reference Model." He emphasized the fact that this was a reference model, not an implementation model. The "open" part of the OAIS refers to the fact that this model was generated in an open forum and is freely available. It does not refer to the level of accessibility of information in an archive. The OAIS framework provides for understanding and applying concepts needed for long-term digital information preservation, with the "long-term" aspect being long enough to be concerned about the changing technologies. This framework could also be a starting point for a model addressing non-digital information. The OAIS provides a set of minimal responsibilities to distinguish an OAIS from other uses of the term "archive." It also serves as a framework for comparing architectures and operations of existing and future archives. On a more practical level, it serves as a basis for development of additional related standards and it addresses a full range of archival functions, such as storage, management, and planning.

Conformance to the OAIS standard by an archive is achieved by discharging the set of minimal responsibilities and supporting the basic information concepts that address a definition of information and types of information packages. A document conforms to the standard by using OAIS terms and concepts. Conformance to OAIS on the part of organizations demonstrates a level of awareness of digital preservation needs. Some definitions followed:

- Information: Any type of knowledge that can be exchanged.
- Data: Representation forms of information.
- Archival information system: Hardware, software, and people who are responsible for the acquisition, preservation, and dissemination of information.

(Continued on page 17)

"Steering by Standards"

(Continued from page 16)

In an OAIS environment, the producer provides the information to be preserved. Management sets the overall OAIS policy and the consumer seeks and acquires the preserved information that is of interest. An OAIS archive must negotiate and accept information from producers, obtain sufficient control over that information to ensure long-term preservation, and determine which communities need to be able to understand the preserved information. It must follow documented policies and procedures to ensure that the information is preserved against all reasonable contingencies and make the preserved information available to the designated communities in forms understandable by those communities.

An information package is a conceptual container holding two types of information: content information and preservation description information (PDI). Deciding what is the content information may not be obvious and may need to be negotiated with the producer. It also may be a digital object or a physical object. The PDI contains different types of information:

- Reference information: Provides one or more identifiers, or systems of identifiers, by which the content information may be uniquely identified (bibliographic description; persistent IDs)
- Provenance information: Describes the source of content information, who has had custody of it, and what its history is (metadata on preservation process)
- Context information: Describes how the content information relates to other information outside the information package (pointers to related collections)
- Fixity information: Protects the content information from undocumented alteration (digital signatures, checksums)

There are variants of the information packages, including the submission information package, the archival information package, and the dissemination information package. Packaging information is information which, either actually or logically, binds and relates the components of the package into an identifiable entity on specific media. Examples of packaging information are: directory structures, filenames, and tape marks. The package description contains the data that serves as the input to documents or applications called access aids. These help the consumer to locate, analyze, retrieve, and order the information. In an OAIS, there are the following functional entities:

- Ingest: Provides services and functions to accept submission information packages from producers and prepare contents for storage and management within the archive.
- Archival storage: Provides services and functions for storage, maintenance, and retrieval of archival information packages.
- Data management: Provides services and functions for populating, maintaining, and accessing (1) descriptive information that identifies and documents archive holdings and (2) internal archive administrative data.
- Administration: Manages the overall operation of the archive system.
- Preservation planning: Monitors the environment of the OAIS and recommends actions to ensure that stored information remains accessible to the designated communities over the long term.
- Access: Helps consumers (1) determine existence, description, location, and availability of stored information and (2) request and receive information products.

In summary, this reference model applies to all digital archives, their producers, and consumers; establishes common terms and concepts for comparing implementations, but does not specify an implementation; identifies a minimum set of responsibilities for an archive to claim it is an OAIS; provides detailed models of both archival functions and archival information; and addresses OAIS information migration and interoperability among OAISs. In conclusion, Sawyer describes some of the follow-on activities and provides some relevant URLs.

(Continued on page 18)

"Steering by Standards"

(Continued from page 17)

Bruce Ambacher next presented "Steering by Standards: The NARA Experience." He began by providing background on the National Archives and Records Administration (NARA) and its tradition of following various standards, as well as participating actively in the development of standards. Ambacher went into some specifics about two NARA programs: the Electronic Records Archives (ERA) and the Archivist's Workbench. The challenge of the ERA is to preserve NARA's electronic records in the face of growing volume, increasing complexity and diversity, and changes in technology. The Archivist's Workbench is studying the scalability of ERA to a smaller archive, in this case, the San Diego Supercomputing Center. In addition to the specific uses of the OAIS in terms of these two projects, Ambacher pointed out other OAIS uses. OAIS can be used to benchmark existing programs. It can illustrate specific program needs and weaknesses and can also help to focus new efforts. It can translate archival program to resource allocators and it can justify targeted improvements.

Last was MacKenzie Smith, who spoke about "Implementing OAIS in Research Libraries." She began by emphasizing the relevance of the OAIS to libraries and archives, who are always acquiring digital material and are mandated to store, manage and preserve that material. This was followed by a description of the Harvard and MIT approaches to creating and maintaining digital libraries. Her use of the common terms from Sawyer's presentation helped to put them into context in the two applications she was describing. Smith's summary of OAIS clearly showed its benefits. OAIS:

- Defines a set of archival functions.
- Defines a detailed information model.
- Maps to the existing digital library infrastructure.
- Is a good model for planning a digital archive.
- Drives related standards development (e.g. METS (Metadata Encoding and Transmission Standard).
- Facilitates communication and interoperation between an archive and its producers and consumers of its information.

In the question and answer session at the end, there were a number of interesting comments. A clear distinction was made between digitization and digital preservation, which are not the same thing. It was acknowledged that in some ways OAIS is redefining archives, so some confusion is natural. But OAIS is a model that should be useful to archives in their organization and planning.

As many conferences do today, both of these raised as many questions as they answered. But it is important to be able to participate in the dialog. So I applaud OCLC for bringing together these knowledgeable speakers on relevant topics and for using technology to include us in the conversation. I am also grateful to the University at Buffalo Libraries for providing access to this series. I look forward to the final videoconference to be presented on May 29th, "Paper Past, Digital Future: Managing Metadata Standards in Transition."

"Steering by Standards"

(Continued from page 6)

The applications are to provide these federated services, synchronize a variety of databases, and to reveal the hidden web, that search engines cannot get to. Data providers may register as being OAI-compliant, though they may use the protocol without doing so. On the service provider side, there is growing interest within the scientific community as well as from web search engine providers. Some institutions are making the OAIMHP freely available and some grant funding institutions are providing funds for use of the protocol.

Van de Sompel discussed the possibility of this becoming a formal standard. The OAIMHP began at the talking stage, then progressed to the current stage where it is viewed as experimental still, but is being practically applied and adjusted. In the future, it may turn out that people will simply accept it as the standard and use it automatically. In answer to a question, Van de Sompel emphasized that a conscious choice was made not to overlap into the areas of search protocols, because others are working in that area. It is simply a harvesting protocol.

Next Stephen Pinfield described the application of the protocol at his institution. He pointed out that their focus was upon institutional e-prints generally. This encompasses much more than just pre-prints. It is basically anything available in electronic format and includes post-review also. Pinfield outlined the relevant issues faced by the University of Nottingham:

- The initial implementation of the OAIMHP is straightforward to set up and provides an "instant" framework.
- There are collection management and development issues that need to be considered. Questions such as: What should be included and in what format? The library staff must deal with these questions, the protocol does not address them. A digital preservation policy is necessary, as are metadata quality standards.
- User participation at all levels must be encouraged through an advocacy campaign. Pinfield said that it became clear the faculty were simply not interested in the serials pricing crisis. It made more sense to play up the incentives for researchers to providing their work in e-print form: higher visibility, rapid dissemination, lower access barriers, value-added services (such as citation counts). An incentive for administrators was the ability to raise the profile of the institution.

Pinfield next focused on some concerns, namely copyright issues, quality control, and workload. He advocated letting the library do the work, but of course it is resource-hungry work, and therefore expensive. Pinfield felt that this work expands the library's role and places it directly in the scholarly communication process. He believes the library should play a major role in managing the resources of the parent institution. There was a specific question about the cataloging workload issue, but Pinfield said that the reorganization was just beginning at the University of Nottingham. The work there is grant-funded and they are pursuing additional grants to keep it going. It is important that cataloging staff be involved in the planning.

Joann Kaczmarek opened her presentation with a question: Can OAI enhance the quality of everyday life? The project she oversees at the University of Illinois at Urbana-Champaign is funded by the Mellon Foundation and is centered around 27 cultural heritage repositories. There are many different data providers involved, including museums, historical societies, libraries, and digital data providers and the focus is on archival communities, rather than scholars. Mapping issues were discussed since the various cultural institutions involved used different metadata, including MARC and EAD (Encoded Archival Description), as well as locally-created.

Kaczmarek repeatedly emphasized the opportunity to expand "discoverability" of these cultural resources. She made the point that unknown collections can be revealed, but so can unknown communities of users. Forming relationships between these unlike resources and users has resulted from this project. Unfortunately expanding discoverability and forming relationships are rather difficult concepts to measure.

(Continued on page 16)