CENDI PRINCIPALS AND ALTERNATES MEETING

Department of Interior, hosted by the U.S. Geological Survey
Washington D.C.
October 2, 2001

Minutes

WELCOME

Dr. Elliot Siegel opened the meeting on behalf of Kent Smith, CENDI Chair, at 9:15 am. He thanked the Department of Interior (DOI) and the U.S. Geological Survey (USGS) for hosting the meeting. Introductions were made. Scott Cameron, Deputy Director for Performance and Management, welcomed CENDI on behalf of the DOI. Mr. Cameron is responsible for driving the President's Management agenda at DOI, including E-government initiatives. His office is particularly interested in interagency groups, such as CENDI, because many of his office's initiatives, including geospatial data for which DOI is the lead agency, are of an interdepartmental nature.

STI POLICY ISSUES FOR FY02

"Perspectives on Information Management and Administration IT Priorities"
Mark Forman, Associate Director for Information Technology and E-Government
Office of Management and Budget (OMB)

Information is at the heart of much of our decision making as a government and as individuals. There will be a fundamental shift from centralized to distributed decision making. CENDI-related information will be key to much of the knowledge that is needed to make critical decisions. A critical question is, "How can agencies support knowledge management for decision making?" Mr. Forman wants to make sure that there is an open dialog with groups like CENDI, and also that the agencies get the right guidance to achieve this goal.

E-government represents a new role for IT in the government. E-government uses the Internet for online filing of forms, to connect to its employees, suppliers and customers, and to transform government operations in order to improve effectiveness, efficiency, and service delivery. The vision for E-government is an order of magnitude improvement in the federal government's value to the citizen. It is an integral part of the principles that the President established in his Management Agenda. Government management practices should be market-based, results-oriented, and citizen-centered. The goal is to simplify and unify.

In terms of federal information systems, there is a special challenge: the beneficiaries and the bill payers are two different groups. It is important to create a value proposition for users. This involves measurable outcomes of quality and service.

The citizen-centered strategy includes four identified segments. One segment is individual citizens for whom the government needs to build easy to use, one-stop shops to high quality government services. FirstGov is a major initiative in this area. It is the primary storefront and a single point of service for business and the public, not just a

search engine. FirstGov is an online service center for government-to-citizen and government-to-business relationships, communication, and transactions. It is moving more to services, including a common services interface.

The business segment is addressed by reducing the burden on businesses through the use of Internet protocols and consolidating redundant reporting requirements. The goal in this and other initiatives is to leverage technologies, such as XML, and to apply best practices in knowledge management.

The Intergovernmental segment includes states and local governments. Again, the emphasis is on reducing the burden of reporting requirements, while enabling better performance measures and better results, especially for grants.

Internal efficiency and effectiveness is also key. This involves reducing costs for government administration by using best practices in areas such as supply chain management, financial management, and knowledge management.

The government has chronic issues related to islands of automation and stovepipes. In the new environment, organizations that have open and interoperable relationships with their customers will have the edge instead of those who keep information to themselves. Optimizing IT spending decisions (based on the Clinger-Cohen Act) will be enforced through the budget process. The communication gap between IT and the lines of business will be addressed by establishing integrated product and process teams. Unfortunately, object-oriented web development has separated the business side from the IT side (including procurement activities).

E-government requires a business case that defines value for the citizen and the government program. Mr. Forman emphasized the requirement for developing business cases. Chances are that a program won't get funding if it doesn't provide one. The components of such a case include the types of constituency groups that are being targeted, the value proposition for these groups, the critical success factors, the organizational structure required, the governance structure for multi-agency initiatives, etc. The value proposition is particularly important since it should be linked to metrics and evaluation. Mr. Forman believes that information management groups, such as CENDI, are vital to the success of these efforts, because information content is central to the development of business cases.

A key component in business case analysis is customer segment assessment. Which types of constituency groups should be targeted? What is the discriminating factor(s)? How should the government provide value to these groups? OMB will provide blanket authority for the CIOs to conduct focus groups for Web sites.

Important goals of the Administration in terms of the government processes are unification and simplification. This involves integrating the organization and the information infrastructure in key business areas. You first have to unify how the government deals with a customer in a line of business, and then simplify the business processes to maximize the benefits of the technologies. There will be seed money from an E-government fund to support change. The E-government Task Force is addressing the barriers to working across agencies.

What will the future look like? It will be a "click and mortar" business design, where the basic requirements of the enterprise are supported by IT. Mr. Forman sees peer-to-peer computing, a sharing model such as Napster, as the key to the future enterprise information management and integration. However, there are issues related to privacy, digital signatures, etc., which are critical since peer-to-peer allows access to the data without having to own it.

Government's role as a content manager and disseminator (publisher) will grow. OMB is seeking to provide content management training within the government. Mr. Forman foresees integrated product teams with CIOs and content managers.

Key governance issues will arise in this environment that go beyond the issue of funding to include executive leadership, business partnerships, ownership of the process and the data, and the relationship with the citizen. A major governance question will be who leads, builds, and controls the integration of the delivery channels.

In support of this vision, Mr. Forman is restructuring the CIO Council. There will be three standing committees: Workforce and Human Capital for IT, Best Practices, and Government-wide Architecture Framework. The latter is likely where Content Management will fit. There will also be one cross-cutting committee, which will probably be security. There will also be four Portfolio Management groups: government to citizen; government to government;

government to business; and internal efficiency and effectiveness (including deploying knowledge management tools and training).

In addition, twenty initiatives, plus two other business cases and the government-wide PKI initiative, will comprise the E-government roadmap. These awards will be announced soon through the E-government web site and the CIO Council's web site.

Discussion

The group discussed the issue of how to share information across the agencies when the information cannot be in the public domain. This arises in particular when coordination for homeland security requires sharing of information across agencies whose only mechanism for doing so is the public Internet. This highlights the need to formalize some of the ad hoc efforts that are going on. We don't know who needs to be involved in the business process or what lines of business should be raised to higher levels.

Mr. Forman noted that he would appreciate support from CENDI for the development of joint data architecture and content models for specific types of information.

PRESERVING FUTURE ACCESS: CURRENT GOVERNMENT INITIATIVES

CENDI has been interested in digital preservation for several years. Its efforts have included joint sponsorship with the International Council for Scientific and Technical Information (ICSTI) of a study on the state of the art and practice in digital archiving. CENDI has a Digital Preservation Task Group, whose members monitor various preservation and archiving activities, both nationally and internationally, and seek to share information and coordinate related activities among the agencies. They have reviewed key documents that other groups have proposed in support of a digital preservation infrastructure. This session is an extension of the group's desire to educate itself and the CENDI members on the efforts of various groups.

"Library of Congress: Developing a National Framework"

Martha Anderson, Library of Congress

The mission of the Library of Congress (LOC) has been and continues to be to make information available and useful, and to sustain and preserve a universal collection of knowledge and creativity – regardless of format. The LOC's digital strategy is an extension of this mission. It seeks to incorporate the digital medium into the unchanging mandate in order to provide access to the record of human experience.

The LOC actually began its digital efforts in 1995 with the National Digital Library Program (NDLP). Preservation naturally came along with the requirements of this program. However, the challenge for the LOC, along with other libraries and archives, is to ensure access to digital materials in an environment of exponential growth and volatile technology. The Web itself is growing at a rate of 250 megabytes per year for every man, woman and child. There are more than 7 million pages added to the Web every day.

In 1999, a study was commissioned from the National Academy of Science called "LC21: A Digital Strategy for the Library of Congress." The results call for the LOC to look outward to others as well as to focus on the library's own internal collection. The report also called for the NDLP to transition from a separate program into the fabric of the LOC itself.

Congress supported this approach late last year when the National Digital Information Infrastructure for Preservation Program (NDIIPP) was established as part of the Consolidated Appropriations Act of FY2001. Funding amounting to $25M was appropriated for initial planning and for acquiring and preserving digital information that might otherwise "be uncollectible". (For example, the LOC recently began capturing Web sites related to the September 11th World Trade Center attacks.) A matching fund of $75M is also included in the appropriations. The deadline for the in-kind or private sector donations is March 31, 2003.

In the development of the NDIIPP, the LOC sees its role as that of facilitator. The LOC is developing a framework for a national information infrastructure for preservation of digital materials through a network of partners, including federal partners and other stakeholder communities. The LOC is looking to its sister national libraries and to other government organizations such as NARA and NTIS to jointly assess planning considerations. OCLC, RLG, and others were specifically identified in the Congressional appropriation as contributors to content, to technology development, or to dissemination of the infrastructure.

Other stakeholder communities will be involved in the development of the infrastructure. These include federal and private libraries and archives; research and business organizations; organizations with expertise in telecommunications, e-commerce, and in the collection and maintenance of digital material; and those involved in efforts to preserve, collect, and disseminate digital information. It is important to begin with those stakeholders who produce digital content, such as e-book and serial publishers, Web-publishers, producers of digital recordings, etc.

Various methods have been identified for actual digital preservation. It is apparent that one size does not fit all, and that there are key differences in archiving approaches based on format, document, and even discipline. NSF is organizing a group to identify a research agenda in this area.

To date, the NDIIPP activities have focused on assembling the National Digital Strategy Advisory Board, which includes representatives of many of the library and archive organizations mentioned above. The LOC is now preparing to convene communities of stakeholders. The management structure is being developed, centered in the Office of the Associate Librarian for Strategic Initiatives.

In addition, a digital strategy framework has been developed. The framework areas include content (including universal holdings), life cycle management of the content, preservation, metadata, technical information, and business policy, rights, and permissions. The LOC is looking to broker licenses for archival purposes and permanent public access.

In early 2002, the proposed plan will be presented to Congress for approval. Approaches will be recommended for developing partnerships, and these partnerships and major projects will be established. The infrastructure will be modeled and tested among the partners.

In 2003, the LOC hopes to achieve the goal of the matching $75 million in funds. The work will be evaluated and recommendations for long-term implementation will be made to Congress.

"Government Printing Office and OCLC: A Partnership for Preservation"
George Barnum, Government Printing Office

The Federal Depository Library Program (FDLP) involves 1,300 Congressionally designated libraries that have been designated as government repositories. This program has existed for 105 years under current statute. Issues of permanence came later to the program than access. Originally, all the libraries got everything, and they could discard the materials as needed. However, in the 1960's, an element of permanent access was incorporated. Fifty-two of the libraries were designated as permanent repositories. They receive everything and have agreed to keep it forever. The others can be selective. Five years after receipt they can begin a discard process.

The FDLP is currently at the end of a multiyear transition to a more (mostly) electronic program. In 1996, Congress directed the FDLP to move to being totally electronic within two years, but the schedule was modified to a more achievable four- to five-year plan.

The goal is to replicate what worked in the old paper model while integrating technology and making changes that are effective. In 1996, they added new locators to facilitate access. In 1998, they chose PURLs as the persistent identifier technology, because that was what was available at the time. This was a quiet but important step toward digital preservation. The FDLP electronic collection policy was written in 1998. It requires each library to also have a written policy.

The goal is a comprehensive digital library of government electronic publications. This goal results in several challenges. For example, GPO's relationships with the libraries were instantly altered when the FDLP went electronic. In the paper environment, there was no centralized responsibility on the part of GPO after the documents had been distributed to the FDLP. Now, GPO is taking a more centralized approach to distribution with fewer physical locations for the content and more distributed access. This peer-to-peer model has run counter to the way that the Internet works. Paper and microfiche distribution continues but it declines every month. GPO has a multitude of legacy systems. Routinely, 12 systems are used to produce and distribute a government document. There is little or no integration between the systems, except that imposed by knowledgeable staff.

After developing the electronic program for the FDLP, GPO found that electronic publications are only as permanent as each agency decides they will be. Therefore, GPO determined that it needed a strategy for preservation and long-term access. They began to ask questions; such as, what does it mean to be added to the electronic collection. They determined that it means that GPO has found the document and has brought it under bibliographic control. Archiving is a natural extension of this definition.

However, GPO found that there is no existing solution to archiving that answers all needs. Therefore, they are using multiple approaches, including agreements with agencies, partnerships with FDLP members, in-house archives, and contractor arrangements.

The work with Online Computer Library Center (OCLC) began with a pilot project between NLE, GPO and OCLC to archive ERIC documents. After the project ended, they continued discussions with OCLC. The key features of the system being developed include tools for discovering, documenting, and describing digital resources. It involves the integration and modification of existing OCLC and GPO systems over time.

The system will allow for multiple repositories, not just those at OCLC or in-house at GPO. It provides for uniform preservation and structural metadata that will coexist with the bibliographic metadata. Preservation metadata has been a hot topic worldwide, and a number of prototypes have resulted, including those in the Netherlands, the UK, and Australia. OCLC has been comparing the various approaches to see what preservation elements appear in most of them. OCLC will then prioritize a set of elements to implement.

The current tools are built on OCLC's CORC (Cooperative Online Resources Cataloging) Project. It also integrates archive management, persistent locators and bibliographic control.

The first phase is built on the CORC web interface for harvesting and describing digital objects. The preservation metadata has been added, along with the link to the bibliographic data creation. A skeletal bibliographic record is created and the preservation metadata is linked to it.

The second phase will add automated discovery and harvesting. It will also implement the OCLC maintained repository. In subsequent phases, the project will deal with managing more complex documents and documents that aren't linear but dynamic.

When GPO looked at its 12 legacy systems, they determined that it was all about metadata. The new system will gather data from existing processes and track the publications throughout the processing cycle. The current processes are very industrial in nature. The goal is to route the publications to the archive through a more automated workflow. The system will eventually assist in discovery and unify the bibliographic description with the acquisitions and classification processes.

The first pilot phase began in July and should conclude in December 2001. Training of staff in the cataloging tools is about 50 percent complete. Catalogers have entered approximately 50-100 "globs" of metadata. (GPO is trying to avoid the use of the word "record".) Functionality is very rudimentary right now. There are 16 metadata elements that are being stored for preservation purposes. The second pilot phase will include discovery and capture as well as routing of material to the OCLC vault. There will be between 16 and 23 metadata elements in this phase.

"National Archives: Handling Petabytes"

Ken Thibodeau, National Archives and Records Administration

In 1998, the National Archives and Records Administration (NARA) did a quick scan and determined that it received its first electronic record in 1971. By 1998, it had approximately 60,000 electronic files. NARA has received approximately 1 million diplomatic messages from the Department of State alone. At the end of the Clinton Administration, 40 million e-mail messages were transferred from the White House to NARA. NARA determined early on that technology could not be scaled up to handle even 10 million files.

When NARA began its Electronic Records Archive Project, there was no proven way to preserve demonstrably authentic records. It is necessary that the solutions be dynamic because technologies will continue to change. In addition, NARA wanted to continue to enhance customer service and to take advantage of using today's tools for yesterday's content. However, the archival community was (and remains) too small to drive technological advances to address these needs, even though the archival community has the most technically complex issues. Therefore, NARA decided that it needed to look for solutions that could be implemented as general government IT, but that could handle the volume and size of digital collections. The Electronic Records Archive (ERA) is one of the Research and Exploratory Development projects created to solve critical problems.

The initial work on the ERA was performed by the University of California at San Diego (UC-SD) under contract to NARA and the US Patent and Trademark Office (PTO). PTO is particularly interested in integrating records management with the process of archiving. Patent information is also extremely complex. PTO actually has a patent application and validation authoring tool in place. UC-SD continues as the ERA contractor.

The ERA infrastructure is based on gigabit-per-second speeds, peer-to-peer computing, and distributed processing. Distribution and redundant storage are considered important. From the policy standpoint, the infrastructure must provide access and, at the same time, consider statutes that restrict access to certain materials for specified periods of time. Thirty percent of the usage of archival material is by the federal government itself.

The architecture is very modular, with the goal of being able to replace any of the hardware and software without changing the system itself. Three levels are addressed: data, information and domain-related knowledge. Mediator software called the Storage Resource Broker integrates the functions and the software. This keeps the system infrastructure independent, so that the technology can be changed with minimal impact on the authentic records.

The properties of the objects to be archived are expressed in formal models, using XML and topic maps. This makes them self-describing and self-validating. The objects are wrapped in the metadata. A key requirement for authentic archives is to be able to rebuild the collection, since each record is defined by its collection and by the context of the record; for example, time sequence.

ERA is based on the Open Archival Information System (OAIS) Reference Model. NARA is involved in developing the criteria for self-certification of an OAIS-compliant archive. NARA is also developing the Archivist's Workbench, which includes tools for accessioning, archiving and reference. OAIS functions are the basis for the design of this tool.

While the goal of the ERA is to deal with the broad spectrum of digital object types, NARA has attacked the problem by class of digital object. Each year they have added one or more object classes for the contractor to address. To date, they have not found any type of material for which the approach does not work. In addition, they recognize that context is important, particularly in an archive. XML schema are used so that the descriptions can be standardized without dictating the structure. The translator will take the content of the repository and represent it on an arbitrary technology at the time of access. XSL is used for instantiation. This is beyond the current state of the art for non-textual objects, but is getting closer.

It is now important for stakeholders to become part of the digital records preservation process. NARA is to the point of being able to give advice to agencies that are seeking to address the archiving of digital material. Several guidance documents are available from the NARA web site.

In addition to the ERA, NARA participates in international digital archive projects. InterPARES involves 13 countries and is multidisciplinary. The preservation task force of InterPARES has developed a model based on OAIS, which tells

what the processes are that must be in place to be a compliant archive. The framework will be released in January. Current work involves translating the model into local instantiations.

Discussion

Several members raised the issue of selection criteria -- not everything that is born digital can be preserved and not everything is worth saving. What criteria should be used to make this decision? NLM has developed a scheme for coding the retention of its digital materials, particularly web sites.

Another key issue is the sustainability of any approach that is taken. Many projects are trying to determine what the long-term costs will be. In the past, an endowment model has been discussed as a mechanism for funding ongoing preservation activities. OCLC is seeking to build preservation into the organization structure as an ongoing function. Congress is also concerned about sustainability and is looking to the NDIIPP to gather information to support decisions in this area.

Ms. Hodge (CENDI Secretariat) and Eleanor Frierson (NAL) were present at the NSF Long Term Preservation Meeting mentioned by Ms. Anderson. The group identified need areas ranging from migration and storage technologies to the establishment of appropriate roles for the various stakeholders in the information life-cycle management. Subsequent to the meeting, several topic areas were identified as benefiting most from an NSF-interagency basic research initiative, many of which center on issues such as the management of large volumes of information, and the digital file characteristics that impact preservation. The group has decided that a workshop is warranted to further identify the research agenda. The group will meet in November to develop a workshop plan.

Return to Minutes Archive