

Notes from NDIIPP meeting with federal sector NDSAB representatives

Discussion of standards, practices and guidelines

Members of the federal agencies that are part of the National Digital Strategy Advisory Board met May 5 in Washington to discuss the standards used in the creation and archiving of digital materials.

The meeting at the Sofitel Lafayette Square hotel brought together representatives of the federal agencies listed in the National Digital Information Infrastructure and Preservation Program (NDIIPP) legislation and members of the National Digital Strategy Advisory Board. The attendees included agency executives and technical specialists: Mary Brady, National Institute of Standards and Technology; Margaret Byrnes, National Library of Medicine; Robert Chaddock, National Archives and Records Administration; Chris Cole, National Agricultural Library; Cindy Etkin, U.S. Government Printing Office; Stephen M. Griffin, National Science Foundation; Betsy Humphreys, National Library of Medicine; Michael Kurtz, National Archives and Records Administration; James Mauldin, U.S. Government Printing Office; Dianne McCutcheon, National Library of Medicine; Maria Pisa, National Agricultural Library; Steve Puglia, National Archives and Records Administration; Judith C. Russell, U.S. Government Printing Office; George V. Smith, Institute of Museum and Library Services; and Peter Young, National Agricultural Library.

Laura Campbell, associate librarian for Strategic Initiatives; Elizabeth Dulabahn, senior adviser for integration management; and program managers Carl Fleischhauer, Babak Hamidzadeh, Kevin Novak and Michael Stelmach represented the Library of Congress.

The meeting began with the representatives highlighting key digital-preservation-related activities. Among the items discussed was the Document Type Definition for e-Journals promulgated by the National Library of Medicine and recently endorsed by the Library of Congress and the British Library (see [news release](#)). The existence of a specialized forensic

registry of software at the National Institute of Standards and Technology was also highlighted. This registry provides a means for the identification of certain types of digital documents. Representatives from the National Archives and Records Administration described a risk-assessment strategy they were developing to identify endangered digital content, which would then be appraised and archived. Library of Congress staff noted the relationship between the use of standard data formats and the development of automated process controls, an approach that will increase the efficiency of a digitizing production line.

Three agencies described their development of repositories for digital content preservation. The National Archives and Records Administration provided an update on the Electronic Records Archives project. The Government Printing Office described its repository effort, noting how standardization came into play in the definition of a submission package and methods for document authentication. Staff from the National Agricultural Library described the agency's repository development project, which is associated with the U.S. Department of Health and Human Services Office of Rural Health Policy.

Policy matters that cut across several government agencies were discussed. GPO representatives noted that they have used the CENDI (Commerce, Energy, NASA, Defense Information Managers Group) as a forum to articulate the role of digital copies as preservation copies, supplanting older guidelines that favor microforms. Another policy-related topic for this group concerned the ways in which textual information may be structured in government records and documents. The participants noted the widespread and appropriate use of PDF and also called attention to the value of markup-language formatting (e.g., XML) for long-term preservation management. Marked-up texts, some participants said, have special value when a large number of documents have been preserved and are managed as a combined textual resource.

The participants also noted the effects of recent actions at the National Institutes of Health that require public access to tax-supported scientific and medical research. The importance of such policies was reinforced by comments from the National Agricultural Library, which has encountered instances in which Department of Agriculture staff have placed articles with

commercial publishers under agreements that limit the agency itself from disseminating the information. Regarding work by recipients of [NDIIPP-National Science Foundation grants](#), there was discussion of the potential value in requiring future grantees to include content preservation in their project plans and budgets. The general topic of preserving scientific data was proposed for discussion with the National Coordination Office for Networking and Information Technology Research and Development, part of the White House Office of Science and Technology Policy.

The group offered some insights into at-risk digital content associated with their agencies. GPO staff cited the prevalence of government information on CD-ROM and DVD, noting that the rapid obsolescence of the software resulted in the disks becoming unplayable. Several agencies, including the National Technical Information Service, reported that it was difficult to identify content that may exist, a significant portion of which is presented to the public via the Web in semiformal ways. Other participants noted the phenomenon of document removal from the Web by agencies of federal and state governments, often when a new administration takes office. Several attendees reported that some databases and scientific datasets were not being preserved. Many of these datasets underpin or are important corollaries to published reports. In a related matter, and presented as a partial solution, National Library of Medicine staff reported that they were working on guidelines for incorporating supplemental materials for articles. The National Agricultural Library is building a repository for “gray literature,” i.e., unpublished or in-draft works, from the field of agriculture.

The central issue of metadata and related standards received attention from the group. Content description supports search and retrieval while technical metadata supports preservation management. Metadata is also central to each agency’s workflow and is a key element in the efficient and effective movement of content through its life cycle. The archivists who were present called attention to standards pertaining to records management, including specifications in use at the National Archives and the Department of Defense. A related matter concerned the deletion of records. Attendees pointed out that retaining metadata about a deleted item would

serve to inform the public about the document's prior availability, thereby maintaining the transparency of government processes.

At the meeting's end, the group drafted a working checklist of possible next actions and proposed that two or three high priority be taken up at the next meeting.

1. Review and complete the draft inventory of standards and practices handed out at the meeting. Using this document as a starting point, develop a gap analysis that identifies needed standards, practices, and guideline statements.
2. Develop a policy statement regarding the deletion of documents or other data from a resource. The statement should consider the need for disclosure of actions taken, e.g., the annotation of and continued access to relevant metadata.
3. Draft a statement that supports appropriate access to federally funded data, consistent with recent policy statements from the National Institutes of Health, and reaffirm that work by federal employees falls in the public domain, and ought also to be openly accessible. Consider addressing this topic with the National Coordination Office.
4. Draft a collective agency position on when digitization constitutes preservation. (High priority: GPO will send their draft position statement to the group)
5. With NSF and the [National Coordination Office](#), explore research/granting language to require a basic level of data preservation in conjunction with awards. (High priority)
6. Coordinate on criteria and procedures that pertain to the harvesting and preservation of Web sites. (High priority)
7. Consider and discuss such topics as
 - a. the value of markup-language encoding for the preservation of textual content;
 - b. issues pertaining to the preparation of formatting of government information that may be presented on CD-ROM or DVD disks;
 - c. interagency planning for the preservation of scientific and other datasets.