Report on the Meeting of Experts on Digital Preservation: Metadata Specifications

U.S. Government Printing Office
Washington, D.C.
June 14, 2004

I. Introduction and Purpose of the Meeting

On Monday, June 14, 2004, Judy Russell, Superintendent of Documents and Managing Director, Information Dissemination, at the United States Government Printing Office (GPO) welcomed a select group of experts on digital preservation. This is the second of a series of meetings with digital experts. In March 2004, GPO held a meeting that focused on digital specifications for digital preservation masters. In that meeting, it was decided that an additional meeting focusing on metadata issues was needed. The Report of the Meeting of Experts on Digital Preservation: Digital Preservation Masters is located at: <http://www.gpoaccess.gov/about/reports/preservation.html>.

The goal of the meeting was to look at issues associated with the development of metadata specifications and to develop a plan to determine a recommended set of specifications for both descriptive and preservation metadata for the digitization of the historical Federal Depository Library Program (FDLP) collection. In conveying that goal, it was also noted that there are a number of institutions, some of which were represented at the meeting, that are ready to begin digitization projects and are waiting for GPO to

declare preferred specifications so that those institutions will be in harmony with GPO's plan for digitization of the historical collections held in the depository libraries.

The goal was not to arrive upon a decision that day, but to compile the participants' ideas and recommendations. Then GPO would promulgate the recommendations and seek comments prior to making any final decisions about metadata specifications for the digitization project for the historical collection.

II. Attendees

Participants in the Meeting:

Linda Cantara.....Case Western Reserve University
Martha Fishel.....U.S. National Library of Medicine
Matthew Gibson.....University of Virginia
Ann Green.....Yale University
Rebecca Guenther.....Library of Congress
Cathy Hartman.....University of North Texas
Gail Hodge.....CENDI[1]
Nancy Hoebelheinrich .....Stanford University
Irene Kavalek.....U.S. Geological Survey
Pamela Mason.....U.S. National Archives and Records Administration
Betty Meagher.....University of Denver
Julie Schwartz.....Connecticut State Library
David Seaman.....Digital Library Federation[2]
Brian Tingle.....California Digital Library[3]
Edward Van Gemert .....University of Wisconsin
Robin Wendler.....Harvard University
Elaine Westbrooks .....Cornell University
Perry Willett.....University of Michigan

---

[1] CENDI is an interagency working group of senior Scientific and Technical Information (STI) Managers from 12 U.S. federal agencies. CENDI's mission is to help improve the productivity of federal science- and technology-based programs through effective scientific, technical, and related information-support systems. In fulfilling its mission, CENDI agencies play an important role in addressing science- and technology-based national priorities and strengthening U.S. competitiveness.

http://www.dtic.mil/cendi/index.html

[2] The Digital Library Federation's mission is to bring together -- from across the nation and beyond -- digitized materials that will be made accessible to students, scholars, and citizens everywhere, and that document the building and dynamics of America's heritage and cultures. The DLF partners with over 30 prominent institutions.
http://www.diglib.org/dlfhomepage.htm

[3] The California Digital Library is the University of California's 11th University library. It was established in 1997 by University of California President Emeritus Richard Atkinson to build the University's digital library, assist campus libraries with sharing their resources and holdings more effectively, and provide leadership in applying technology to the development of library collections and services.
http://www.cdlib.org/

Observers:

Prudence Adler.....Association of Research Libraries
Beth Camden.....University of Virginia
Martha Crawley .....U.S. Institute of Museum and Library Services
William LeFurgy.....Library of Congress
Clifford Lynch .....Coalition for Networked Information
Barbara Paulson .....U.S. National Endowment for Humanities

GPO Representatives:

Gil Baldwin
T.C. Evans
Laurie Beyer Hall
Robin Haun-Mohamed
Yvonne Louden
Judy C. Russell
Kelly Seifert
Scott Stovall
Mike Wash

III. Summary of Discussion

Introductory Comments

The meeting began with a welcome to the participants and an introduction to the GPO preservation initiatives by Judy Russell. To put GPO's preservation initiatives into context, Judy Russell quoted Bruce James, the Public Printer of the United States, saying,

> The U.S. Government Printing Office's core mission, Keeping America Informed, dates to 1813 when Congress determined the need to make information regarding the work of the three branches of government available to all Americans. This is the inherent function of government which GPO carries out for Federal agencies on behalf of the public. The GPO is the Federal government's primary centralized resource for gathering, cataloging, producing, providing and preserving published information in all its forms.
>
> By law and tradition, GPO has been the principal provider of publishing services for the Federal government. There is no other agency with the breadth and depth of skills and the knowledge required for the production and dissemination of published government information in all forms. No other agency is specifically funded by Congress to provide information dissemination services for all branches of the Federal government. The GPO needs to take the lead in creating digital standards for official documents of the United States Government.

> GPO must deploy the technology needed by its federal customers and the public to gather and produce digital documents in a uniformly structured database in order to authenticate documents disseminated over the Internet and to preserve the information for permanent public access.
>
> GPO needs to work with its library partners to develop a new model for no-fee public access through the FDLP, which must include a fully digital database of all past, present and future U.S. Government documents, augmented database search and retrieval tools, and increased training to enable librarians to better serve the 21st century information needs of their patrons. [4]

Judy Russell then explained how GPO is working with the library community to develop a national digitization plan. The main components of this plan are: digitization of the historical legacy collections located in the depository libraries; development of a collection of last resort to ensure permanent public access to the resources of the Federal government; and assistance in the development and implementation of specifications used to develop digital collections.

The historical legacy collections are considered by GPO to be a national treasure that must be preserved and remain available in the public domain for permanent public access. Many of these collections are not being preserved under an active preservation program. Much of the material is deteriorating gradually on the shelves, and action needs to be taken now or in the near future to preserve access to the information for future generations.

Many of the depository libraries no longer have the space or resources to house the historical collections. If the historical publications were digitized, the content would continue to be available for access. Moreover, it would allow those institutions that wish to reduce their collections to do so by substituting electronic access to the digitized copies. The tangible materials could then be moved to a storage facility or weeded from the collection as necessary.

Bruce James then also spoke directly to the participants, briefly discussing how Government printing began with the printing of the Congressional Record and the Congressional journals. GPO is now looking at developing a new Government information system that would gather the documents of Government and organize the documents by uniformly tagging the documents utilizing a standard character set. The system will also allow GPO to easily repurpose that data for printing, for dissemination over the Internet, for making CD-ROMs, or for any new technology in the future.

---

[4] This quote was taken from a presentation given by Bruce R. James, Public Printer of the United States, at the 2004 Spring Federal Depository Library Council Meeting. The presentation was entitled: KEEPING AMERICA INFORMED IN THE 21ST CENTURY: A FIRST LOOK AT THE GPO STRATEGIC PLANNING PROCESS — "A Work in Progress." The presentation is dated May 1, 2004. http://ww1.access.gpo.gov/gpoaccess/fdlp/pubs/proceedings/James.DLC.04192004.revised.pdf

Bruce James also addressed the issues of versioning and authentication. He stated that these are issues of interest that need to be addressed, and that because the Government will not be well served by many disparate metadata standards, it should develop a universal set of flexible metadata standards. This set will be established, not by mandate, but by cooperatively working together to understand the benefits of such a standard.

Bruce James also discussed the need to go back and digitize the historical materials and put them it into a format that will allow the images to be viewed and will allow for character string searching of the data. This will be expensive, but GPO will be working with agency and library partners, with GPO taking a leadership role in the project. In developing this digital collection, GPO will be looking at different technologies to ensure that the materials are available in the future. This includes maintaining tangible copies of publications when printed, and investigating the best preservation formats for born-digital products.

Bruce James concluded with a brief discussion of how the Federal Depository Library Program has been changing for the past ten years, and how he expects the changes to continue, with more emphasis on electronic dissemination and the need to develop tools to assist users in finding information on the Internet. GPO is pursuing the development of such tools, to assist not only those who use *GPO Access*, but also to provide assistance and training for the librarians in the depository libraries.

The Discussions:

The Morning Session: *Descriptive Metadata Schemas Employed by Expert Institutions*

The morning session began with a round-robin session, as participants described their institutions' digital projects and the descriptive metadata currently used in those projects. A detailed spreadsheet of the metadata that each institution utilizes is attached as Appendix A. Participant responses included the following:

- Machine-Readable Cataloging (MARC) Records
- Dublin Core
- Text Encoding Initiative (TEI)
- Encoded Archival Description (EAD)
- Online Information Exchange (ONIX)
- Federal Geographic Data Committee (FGDC) - Content Standard for Digital Geospatial Metadata (CSDG)
- Data Documentation Initiative (DDI)
- Metadata Object Description Schema (MODS)
- Metadata Encoding Transmission Standard (METS)
- Open Archive Initiative (OAI)

Most of the institutions represented used MARC records in association with their digital projects, and many also made use of Dublin Core. There was no clear consensus among

the institutions for metadata schemas other than a reliance on the MARC records. There are many MARC records for the material in the historical collection, and the group generally agreed the best approach is to make use of that information, by starting out with the MARC record at the highest level for descriptive metadata and building on those initial elements.

To further the discussion of the descriptive metadata issues, Judy Russell explained that she wanted to know if the group could determine enough best practices or uniformity of elements that would help to drive GPO toward a preferred descriptive metadata schema. She noted specifically that inventing something would not be ideal. Judy Russell also explained that GPO was looking for a decision that was practical, because GPO has an enormous task to do, and it needs to be done quickly and efficiently, utilizing automated tools whenever possible to minimize the manual processing. She also stated that GPO wants to do the best job possible with the first digitization scan as we may not have a second opportunity to digitize many of the older resources due to their fragile nature and the limited number of copies.

Judy Russell discussed the current trends in federal publishing in that there are now often multiple formats for almost every federal publication. Often there is a print format and in some cases, a microform format, and multiple digital formats. GPO may have the native files (QuarkXPress, Pagemaker, etc.) that we receive from the agency, and we are actively seeking or creating print-on-demand files. GPO intends to create a preservation master from the born-digital products, or at least keeping something that we can go back to. GPO wants to establish one descriptive metadata record to identify and provide access to the same content, whether it is a print object, a print-on-demand PDF file, or a QuarkXPress file.

Several issues of concern were raised about how to develop a metadata schema for the historical collection. The first was the issue of preserving the historical collections of the FDLP. Although the resources are located in depository libraries throughout the United States, it is not a cohesive collection with consistent treatment and maintenance. Each depository collection is incorporated into a larger library collection and managed under the policies and procedures of that institution. Some of the depository is being actively preserved, but by far, this material is generally not being done in a systematic manner. To assist in efforts to preserve these resources, GPO, working with the Center for Research Libraries, has put forth the Decision Framework for Federal Depository Libraries, located at: <http://www.access.gpo.gov/su_docs/fdlp/pubs/decisionmatrix.pdf.>[5]

The group also considered the questions of who will the collection serve, what level of access will be available, and what level of support will be available to access the materials in the collection. The participants in the meeting agreed the answers to these questions are essential to develop an effective digital collection. The group felt very

[5] The April discussion draft was available for meeting participants to review. A revision of the Decision Framework, entitled, Federal Document Repositories: Decision Framework by Tangible Repository Type, dated September 13, 2004, is located at <http://www.access.gpo.gov/su_docs/fdlp/pubs/matrix_repository_type.pdf .>

strongly that the plan and any metadata schema must be developed by GPO—they could provide advice and direction based upon their experience, but the initial plan must come directly from GPO. They would like to see GPO develop a draft plan for digitization of the historical collection project that is similar in scope and coverage at the plan for the National Collection of U.S. Government Publications. This plan, which highlights the use of two dark archives and a light archive to ensure continued preservation of these materials for now and for the near future, is located at: <http://www.gpoaccess.gov/about/reports/clr0604draft.pdf>.

A discussion of the various types of resources found in the historical collections highlighted some of the problems encountered by the visiting experts in their digitization projects. One of the largest problems is the digitization of serials or multi-volume sets. The MARC record does not work well with a one-to-many correspondence between the bibliographic record and the digital objects. A brief exercise with the *Code of Federal Regulations (CFR)* was used as an example of this problem. In the CFR, many titles included in the books contain parts and subparts that are updated each year and yet are not necessarily the same parts and subparts in each book for the next year.

Another area of concern is tabular materials. Information products such as budgets and statistical resources have unique problems associated with digitization. Effective searching of digitized products with nested tables is a problem. One project at Yale, utilizing Mexico's statistical abstracts, focused on making the information in statistical charts and tables available, by moving them from their published context into a dynamic online interactive system. At this time the project has shown this to be an extremely time-consuming and expensive process.

After the mid-morning break, Judy Russell put forth a hypothesis:

> The digitization project will require a MARC record, with TEI headers and Dublin Core records to be created as derivatives from the TEI headers. This will allow libraries to make best use of metadata records that many depository libraries already have in place and allow the scanned material to be exposed to OAI.

The hypothesis was discussed, but not strongly supported by the digital experts. The consensus appeared to be for GPO to focus on developing metadata specifications associated with a wrapper or meta package, which will allow use of different kinds of schema, including MARC records. It was accepted that at the highest level, a MARC record would need to be provided for the digitized items, thus providing a common component from which additional metadata elements can be extracted. METS was raised as an effective solution to these requirements, allowing GPO to utilize the MARC record at the top level and different schemas below, thus providing more flexibility for different levels of metadata to be compiled. A MARC record will provide rich data, and most libraries will be able to contribute a MARC record for the digital objects. For institutions that routinely utilize specialized schema, such as FGDC for geography/biology profiles, this additional information can also be employed. METS allows this flexibility. In addition, information in other resources, such as the ONIX records, would also be

accommodated by METS. But the group expressed the importance of having the tools to enable information sharing—to have a crosswalk from MARC to FGDC or from ONIX to MARC, because these standards exist and they are not going away. There was also a brief discussion of mapping and how the definition differs from institution to institution. Two important problems were identified with crosswalking and mapping: the problem of having comparable metadata expressed multiple times for items with the same information content and the problem with the semantics of mapping relationships among digital objects.

Participants agreed that flexibility was a necessary component of any metadata scheme utilized by GPO for the digitization of the historical collection project. After discussing the problems associated with series/serials and the need for crosswalks, one suggestion put forth by participants seconded by observers was to proceed with a test project with simple tangible items, such as monographs, to allow the development of processes and procedures and to work through the instructions, expectations and specifications partners will need to meet in participating in this digitization project.

The Afternoon Session: Preservation Metadata Employed by Expert Institutions

The afternoon session focused on a discussion of preservation metadata employed by the institutions. Judy Russell explained that the original concept or goal of the meeting was to address the problems of descriptive metadata in the morning session, and the afternoon would be comprised of a discussion of preservation metadata, including the administrative/technical metadata. It was decided to have a round-robin discussion of the elements of preservation metadata employed by the various institutions. Judy Russell then clarified how GPO uses the term preservation metadata as an umbrella term to cover all of the remaining metadata that is not descriptive. In the round-robin discussion it was again apparent that there was no one schema currently used by the participating institutions. Rather, the libraries generally utilized local schema for preservation metadata.

A brief discussion of what the term preservation metadata means followed the round-robin exchange. For some institutions, preservation metadata refers to administrative metadata. For others, it refers to technical metadata, such as the equipment, scanning specifications and other requirements designated by the NISO Data Dictionary – Technical Metadata For Digital Still Images. The limits of digital preservation were also briefly discussed, including the reminder that although the item has been digitized, it does not necessarily mean the data has been preserved—the scans may not be adequate, or over the years, they may not be sufficient to reproduce the object when needed. This allowed the group to briefly reiterate the need for quality control both as the material is digitized and over time as the material is held in storage.

Discussed at length was the effort of the RLG/OCLC Preservation Metadata Implementation Strategies (PREMIS) working group. The group is focusing on developing core elements for preservation metadata, the information needed for long-

term preservation, and technical metadata elements, but only those that are not specific to a file format type. This work builds off of a previous working group's effort of examining OAIS. The group is in the final stages of identifying the core elements, which are expected to be completed by the end of this calendar year.

Digital provenance was discussed in relation to the PREMIS efforts, and also in the general metadata discussion. Provenance, the path of how the publication came into being, what its background is and what is its role in a larger context, is becoming increasingly important, especially as the concept relates to authenticity of a publication. Digital provenance also includes documentation of the transformations of analog materials to digital, digital reformatting information, (e.g., TIF to JPEG, resolution from 600dpi to 100dpi, creation of thumbnails, etc.), and details on who was responsible for what. The need to explicitly document provenance was first raised in the observers' comments session in the morning and reiterated by several experts in the afternoon discussions.

The issues of the different partner libraries who might participate in the project, and the level of support each partner can bring to the digitization project were discussed at length. Are there different acceptable levels of metadata that allow those institutions with fewer resources for digitization at hand to participate in the project? Is it reasonable to ask participating institutions to contribute MARC records for each digital object as part of the metadata requirements? The discussion associated with this issue was similar to the discussion of the first digital experts meeting, when scanning requirements were discussed and a compromise was reached for recommended scanning specifications, which included a lesser specification under some conditions. The metadata experts put forth the need to develop a digitization plan that clearly documents the decision making process for the project, that identifies the requirements for scanning and metadata, and that provides complete details for where and how the digital files will be stored and what will be publicly accessible. Sharing of the metadata specifications will be important not only to those wishing to participate in the process, but also for other institutions considering developing other digital projects.

The group also discussed the importance of a digital registry for preservation masters, such as is currently being developed by DLF, in cooperation with OCLC, to share information among the library community about the materials for which digital preservation masters are available. In addition, the topic of the Global Digital Format Registry (GDFR) was discussed. This registry provides information on formats, which are difficult to acquire or determine, especially after they are no longer in mainstream use. The primary goal of the GDFR is to collect exhaustive and authoritative documentation about all digital formats. The need to establish a reliable resource on format specifications was reinforced during the afternoon discussion.

Another area of concern for the group of digital experts was the issue of the immediate timeframe for the GPO digital initiatives. GPO is moving forward to develop specifications for digital preservation projects, because the projects have already begun, and the sooner the requirements are determined, the more likely the material submitted

will be useful in the national digital collection. Some institutions are holding off on their digital projects, awaiting specifications similar to the scanning standards identified in the March meeting. It became clear in this meeting, however, that the metadata specifications are not going to be as easily identified. Additional work will be needed by GPO to identify the necessary elements, including the review of documents, as recommended by meeting participants. The specific recommendations included the newly issued "Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images," review of metadata standards as established for the National Library of New Zealand, and review of specific institutions' required elements and the subsets developed to deal with their own institutions' needs. GPO has reviewed these resources and information about them is located in Appendix B.

In another exercise to work through the issues associated with digitizing Federal publications, the group looked at the Public Papers of the President, which are currently being digitized by the University of Michigan. The papers go back to the Hoover administration. They are published by the Office of the Federal Register, and include public speeches, veto messages, radio addresses, State of the Union addresses, etc. They contain mostly text, but many have some photos. Any given year of the term may have more than one volume. The discussion started out with a MARC record for the series, which would be the top layer, but there is a need to create additional data for each physical volume. And as there is a volume for the year, and multiple parts for the volume, date coverage for each volume would probably also need to be captured—allowing for navigation between volumes. One question needs to be considered—are multiple records needed for the books, or can the structural metadata meet the need? The answer in part depends on the number of links there are likely to be—for the Public Papers of the President, one record will probably be sufficient, but for the CFR, with hundreds of books, it will probably not work out as well with one record. If the Public Papers of the President are made available via a website, to include not only the print material, but also audio and video clips and still photographs, the material should be broken down to the document level. The consensus of the group was that this title would be a candidate for METS. In addition to the book level breakout, the sequence of pages, the title page, the table of contents, the beginning of the text, and the index would also be useful information to capture. Cornell has developed an in-house image tool to track where images are supposed to be placed in the sequence of the publication. The tool, however, is extremely costly, as it is time intensive, and even if it were available, it would place a large burden on the digitizing institution.

As shown in the above exercise, there are many different levels of participation for this digital project. In a distributed effort there is a level of work that can be done to a certain point, and then additional development is done by another institution. For example, the scanning is done by one institution and the metadata creation is handled by another party. This raised the issue of the value of having an institution participate—is it worth what is contributed? Of concern is the overhead associated with the project—is it cost effective to accept a partial submission, such as volume 1 of a large set, or the digital files, but no accompanying metadata? Also, with a distributed project where the digitization is done separately from the metadata, there is a potential to build up the number of scanned files,

while there is little progress with the development of the metadata. The requirements will need to allow for some lesser participation for institutions that are less well-endowed, but not at the cost of developing bottlenecks or other situations that become logistically impossible.

The experts agree that GPO needs to say to potential digitization partners, this is the minimum of what we need—the sort of scans wanted and what metadata must accompany the scans. The ideal schema will need to be sufficiently flexible to be enhanced or expanded over time. For example, while GPO needs ONIX records, libraries may not have that information. That may be a value added product that GPO brings to the table after the initial information is received. Another example is map records. Some libraries will include FGDC information for maps associated with their state, but for maps associated with other states, will only utilize the MARC record. The schema must allow this level of difference in the metadata records. Guidance will also need to be provided with regard to recording headers, comparing check sums, and developing other metadata requirements, including training for staff as needed. If the potential partners know what is expected up front, they can make decisions on whether or not they can participate.

Conclusions

It became clear from the discussion as the day progressed that in general, the participants endorsed the concept of wrapping the metadata around the digital object, rather than maintaining the metadata in a separate database. And METS continued to be favorably viewed as a possible solution, with its wrapper format that allows diversity of schema at different levels. The participants again discussed how METS allowed for a MARC record at the top level, with Dublin Core, FGDC, and DDI at lower levels, as needed in the wrapper. Participants and observers in general believed this was an effective place to start to develop the metadata requirements for the digitization of the historical collection, thus allowing for maximum flexibility for potential digital partners in the project.

Working as a group, the participants and observers developed a model metadata package with 11 high-level elements that were suggested as required elements for any scanned titles submitted as part of the project to digitize the historical collection. The following is the list of elements:

- a unique title for each digital object (each one of the 2.2 million physical artifacts being digitized)
- MARC record (which may identify the series)
- unique identification number (for the object, the components, and any links back to a parent object)
- the list of files in a given meta package and check sums associated with the files
- structure map
- technical map
- digital provenance

- indication of the level of quality control performed during the digital process
- level of collection administration
- information about derivatives
- information about property rights

Although the purpose of the meeting was to set forth for consideration metadata specifications for depository libraries to follow in digitizing objects in the legacy collection, it was determined that additional effort was needed by GPO before such specifications could be put forth. Additional review was done of existing metadata specifications, such as those of the National Library of New Zealand Metadata Framework, NARA's recently released "Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images," and specific subsets of metadata elements of already established standards (see Appendix B). These are already in use in local schema at some of the participating institutions represented at this meeting. GPO agreed to prepare a summary of this meeting and a set of metadata elements as developed after reviewing these additional resources. These specifications were then sent to the digital experts for review and response in accordance with the directions provided in this meeting. See Appendix C for the listing of metadata elements developed in response to the advice provided by the experts in this meeting.

Appendix A:

Descriptive Metadata Schemas Employed

Institutions	Data Documentation Initiative	Dublin Core	Encoded Archival Description	Content Standard for Digital Geospacial Metadata	Machine Readable Cataloging	Metadata Encoding Transmission Standard	Metadata Object Description Schema	Open Archives Initiative	Online Information Exchange	Text Encoding Initiative Headers
Case Western Reserve University		X			X	X	X			X
California Digital Library		X	X			X	X		X	X
CENDI		X	X	X	X			X		
Connecticut State Library		X			X	X				
Cornell University		X		X	X					
Digital Library Federation						X		X		X
Harvard University	X		X	X	X	X	X			
Interagency Committee on Government Information		X								
Library of Congress		X	X		X	X	X			
U.S. National Library of Medicine					X					
Stanford University		X	X		X	X				X
University of Denver		X								
University of Michigan	X	X			X					X
University of North Texas		X			X					
University of Virginia		X			X					X
University of Wisconsin	X	X			X					
United States Geological Survey		X		X	X					
Yale University	X	X	X		X					

Appendix B:

Post Meeting Review of Additional Resources

Review of NARA's Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files - Raster Images, provided additional references to the PREMIS work and the OAIS reference model. But as the guidelines focus on the process of digitizing archival material for electronic access, not the long-term management and preservation of digital masters, there is limited application to the development of preservation metadata specifications for the digitization of the historical collections in depository libraries. The revised Technical Guidelines are located at <http://www.archives.gov/research_room/arc/arc_info/guidelines_for_digitizing_archival_materials.html>.

The National Library of Australia has developed a preservation metadata set that supports the management of a digital collection. It is meant to be a data output model—it indicates the information the NLA wishes to be able to obtain from the metadata system. The model does not focus on what data should be entered, how it should be entered, by whom and at what time; nor does it concern itself with how the metadata should be associated with what it is describing. According to NLA, "This model simply says: 'however you do it, this is what you have to deliver so we can manage preservation.'" It consists of 25 elements, with many sub-elements and repeatable fields. Information about the model is located at <http://www.nla.gov.au/preserve/pmeta.html>.

The National Library of New Zealand has also developed a preservation metadata model. The data fields are based on the following structure: Object, Process, File and Metadata Modification. The data dictionary defines the preservation metadata fields from a design or implementation perspective. The dictionary is based on the logical preservation metadata model and maintains the overall structure and data relationships contained there. Individual fields however have been adjusted to facilitate their population with readily available structured data that conforms to recognized standards. In some instances this has resulted in the logically defined fields being split into subfields to enable multiple dimensions or sub elements to be recorded. The four entities are object, process, file, and metadata modification and each allow for additional subfields.

A listing of Metadata Elements/Qualifier Display Names currently in use by Stanford University was provided by Nancy Hoebelheinrich to GPO after the meeting. This subset of 57 required elements is based on NISO Z39.87-2002 AIIM 20-2002 Data Dictionary – Technical Metadata for Digital Still Images. In addition, Stanford recommends the inclusion of 19 additional elements, including subject and coverage, but these remain recommendations, not mandatory elements. Review of this element subset confirms that a workflow model and the required elements must be in place before the digitization project begins to ensure sufficient information is available for preserving access to the digital object.

The University of North Texas has developed a draft metadata model to provide access to the information product. The approach taken is one that will minimize the risk of digital resources from becoming inaccessible. The metadata needs to be consistently maintained throughout the process, best achieved when the metadata is done in a consistent and uniform manner. The detailed workflow and user guide document provides procedural information required to create metadata with examples for different file formats. Because the metadata assigned to an item entirely depends on the metadata creators' definition of the work, the detailed user guide is necessary to provide rules, syntax and descriptive information to identify the source of information for each of the 27 elements.

The Categorization of Government Information (CGI) Working Group, one of three working groups that comprise the Interagency committee on Government Information (ICGI), recently released a document, "Common Characteristics for Government Information Resources," dated August 19, 2004. The document represents the current state of work from an ad hoc group that is working to identify common elements for metadata. Both the Web Content Management Working Group and the Electronic Records Policy Working Group recommended data elements and the ad hoc group reviewed and recommended 13 elements as a set of metadata. With the exception of two, all elements are mandatory and the set of metadata is considered incomplete unless all the mandatory elements have been included. The 13 elements are: access constraints, copyright, creator, date created, date reviewed, description, disposition authority, event, identifier, language, line of business, title, vital records indicator. Further information about the ICGI and CGI working group can be found at <http://www.gpoaccess.gov/cgiwg/ >.

In reviewing the literature shown above, it is clear that the best practice is to create the metadata at the information creation stage—this is where the long-term archiving and preservation must start. Metadata routinely collected at this point would be relatively easy, consistent, reliable, and automatic. However, much of the preservation metadata continues to be created in a manual way, usually after the material has been digitized. Metadata created in this way does not allow for the easy creation of the elements' records. Additional resources must be utilized to ensure record creation is done consistently and in accordance with the metadata plan. Standards groups and others interested in developing consistent and useful metadata schemes continue to work with industry officials and other interested parties to incorporate XML and RDF into their word processing and other software products, thus, providing for the creation of metadata as part of the origination of the object.

Appendix C:

List of Metadata Elements

Meta Element #	Element #	Element Display Name	Attribute Display Name	List of Control Values	Mandatory
		DESCRIPTIVE			
1	D-1	RESOURCE_ID			Y
2	D-1.1		RID_TYPE	ISBN, LCCN, ISSN, DOI, other	
3	D-2	CREATOR			Y
4	D-2.1		NAME_TYPE	personal, corporate, service, organization	
5	D-2.2		CREATOR_ROLE	author, performer, photographer, editor, other	
6	D-3	TITLE			Y
7	D-3.1		TITLE_TYPE	collection, main, alternative	
8	D-4	RESOURCE_TYPE		collection, dataset, event, image, interactive resources, service, software, sound, text	Y
9	D-5	DESCRIPTION			Y
10	D-5.1		DESCRIP_SOURCE	TOC, abstract, related reference, doc analysis	
11	D-6	SUBJECT		Y	Y
12	D-6.1		SUBJECT_TYPE	Y	
13	D-6.2		SUBJECT_SCHEMA	Y	
14	D-7	COVERAGE		N	Y
15	D-7.1		COVERAGE TYPE	Y	
16	D-7.2		BEGINDATE	N	
17	D-7.3		ENDDATE	N	
18	D-8	DIGI_RESOURCE_DATE			Y
19	D-9	CONTRIBUTOR			Y
20	D-9.1		CONTRIBUTOR_ROLE	donor, distributor, sponsor, service, sanning agency, issuing unit	
21	D-10	DIGITAL_PUBLISHER			Y
22	D-10.1		DIGPUB_NAME_TYPE	personal, corporate, service, organization	Y
23	D-10.2		DIGIPUB_ROLE	scanning agency, encoding agency, publisher, provider, vendor, aggregator	Y
24	D-11	LANGUAGE		eng, fre, spa	Y
25	D-12	RELATED_RESOURCE			Y
26	D-12.1		REL_URI_TYPE	URL, Unicorn catalog key, ISBN, ISSN, DOI, SICI	Y
27	D-13	SOURCE			
28	D-13.1		SOURCE_PUB_NAME		
29	D-13.2		SOURCE_PUB_DATE		
30	D-13.3		SOURCE_PUB_PLACE		
31	D-14	RIGHTS STATEMENT			Y
32	D-15	ACCESS CONSTRAINTS			Y
33	D-16	ACCESS FACILITATORS		SYSTEM/METHOD USED TO ENAHNCE ACCESS-NEED TO BE MAINTAINED IN SUCCESSIVE GENERATIONS	
34	D-17	AUDIENCE			Y
35	D-18	LAST REVIEW DATE			Y
36	D-19	KEYWORDS			Y
37	D-20	HOLDING INSTITUTION			Y
38	D-21	ORGINAL DATE			
39	D-22	OBJECT COMPOSITION		file types that make up object content	
		SOURCE			
40	S-1	SOURCE_TYPE		born digital, book archival resource, manuscript, journal, map, photograph, microform, video frame, audio recording	Y
41	S-2	TRACKING_VALUE			Y
42	S-2.1		TRACKING_TYPE	barcode, accession number	Y
		TECHNICAL FOR EACH FORMAT			
43	TA-1	CHECKSUM			Y
44	TA-1.1		CHECKSUM_TYPE		Y
45	TA-1.2		CHECKSUM_VALUE		Y
46	TA-1.3		CHECKSUM_DATE_TIME		Y

Meta Element #	Element #	Element Display Name	Attribute Display Name	List of Control Values	Mandatory
47	TA-2	EVENT		tiff creation, plain text creation, XML encoding, PDF creation	Y
48	TA-2.1		EVENT_TYPE		Y
49	TA-2.2		DATE_TYPE_CREATE		Y
50	TA-2.3		DATE_TYPE_MODIFY		
51	TA-2.4		DATE_TYPE_ACCESS		
52	TA-3	OBJECT_FILE_SIZE			Y
53	TA-4	FILE_NAME			Y
54	TA-5	FILE_PATH			Y
55	TA-6	MIMETYPE		image/tiff, image/jpg, application/pdf, image/gif, text/plain, text/xml	Y
56	TA-7	BYTE_ORDER		big, little, middle	Y
		TECHNICAL FOR STILL IMAGES			
57	TI-1	COMPRESSION_SCHEME		uncompressed, CCITT 1D, CCITT Group 3, CCITT Group 4, LZW, JPEG, PackBits	Y
58	TI-2	COMPRESSION_LEVEL			Y
59	TI-3	COLORSPACE		WhiteisZero, Blackis Zero, RGB, palette color, transparency mask, CMYK, YCbCr, CIELab	Y
60	TI-6	ICC_PROFILE_NAME			Y
61	TI-6.1		ICC_PROFILE_URL		Y
62	TI-6.5		REFERENCE_BLACK_WHITE		Y
63	TI-7	SEGMENT_TYPE		strips, tiles	Y
64	TI-7.1		STRIP_OFFSETS		Y
65	TI 7.2		ROWS_PER_STRIP		Y
66	TI-7.3		STRIP_BYTE_COUNTS		Y
67	TI-8	PLANAR_CONFIGURATION		chunky, planar format	Y
68	T-9	IMAGE_FILE_SIZE			Y
69	TI-10	ORIENTATION		normal, normal rotated 180 degrees, normal rotated cs 90 deg, normal rotated ccw 90 deg, unknown	Y
70	TI-11	IMAGE_SOURCE_TYPE		daguerrotype, reflection print, silver gelatin print, Acme Bronze 100, chroagenic film, color negative, microfiche, microfilm	
71	TI-13	OS		Windows, Mac, Unix, Linus	Y
72	TI-14	OS_VERSION			Y
73	TI-15	DEVICE_SOURCE		transmission scanner, reflection print scanner, digital still camera, still from video	Y
74	TI-16	SCAN_MAKER			Y
75	TI-17	SCANNER_MODEL			Y
76	TI-18	SCANNER_MODEL_NO			Y
77	TI-20	SCAN_SW			Y
78	TI-21	SCAN_SW_VERSION			Y
79	TI-22	X_PHYS_SCAN_RESOLUTION			Y
80	TI-23	Y_PHYS_SCAN_RESOLUTION			Y
81	TI-24	METHODOLOGY			
82	TI-25	SAMLE_FREQ_PLANE		camera/scanner, focal plane, object plane, source object plan	Y
83	TI-26	SAMPLE_FREQ_UNIT		no absolut unit, inches, centimeter	Y
84	TI-27	X_SAMPLE_FREQ			Y
85	TI-28	Y_SAMPLE_FREQ			Y
86	TI-29	IMAGE_WIDTH			Y
87	TI-30	IMAGE_LENGTH			Y
88	TI-31	BITS_PER_SAMPLE		1;4;8; 8,8,8; 16,16,16; 8,8,8,8;	Y
89	TI-32	SAMPLES_PER_PIXEL		1,3,4	Y
		TECHNICAL FOR TEXT			
90	TT-1	OCR_ENCODING_QUALITY			Y
91	TT-2	ENCODING_HW_PLATFORM			Y
92	TT-3	ENCODING_SW_PLATFORM			Y
93	TT-3.1		ENCODING_SW_VERSION		Y
94	TT-4	ENCODING AGENT		ocr,transcriber, markup, editor	Y
95	TT-4.1		ROLE	editor, ANSI X3.4-1968, ISO 8859-1	Y
96	TT-5	CHARACTER SET		Latin 1	Y

Meta Element #	Element #	Element Display Name	Attribute Display Name	List of Control Values	Mandatory
97	TT-6	LINEBREAK		CR, CR/LF	Y
98	TT-7	MARKUP_METALANGUAGE		SGML, XML, GML	Y
99	TI-7.1		MARKUP_METALANG_VERSION	TEILite June, 1995 with May 2002 revisions	Y
100	TI-8	MARKUP_LANGUAGE			Y
101	TI-8.1		MARKUP_LANGUAGE_VERSION		Y
102	TI-10	LANGUAGE		subset of ISO 639-2 code list, e.g., eng(English); fon (French), nbl (Spanish)	Y
		TECHNICAL FOR APPLICATION			
103	TAP-1	APPLICATION_NAME		PDF, MSWORD, MSEXCEL	Y
104	TAP-2	APP_CREATION_SW			Y
105	TAP-2.1		SW_ROLE	Y	Y
106	TAP-3	APP_CREATION_SW VERSION			Y
107	TAP-4	APP_SPECIFICATION		PDF Level 1.2, PDF Level 1.3, PDF Level 1.4	Y