Implementing Preservation Repositories For Digital Materials: Current Practice And Emerging Trends In The Cultural Heritage Community

_____

A Report by the PREMIS Working Group

September 2004

_____

PREMIS—Preservation Metadata: Implementation Strategies

A Working Group Jointly Sponsored By OCLC and RLG

Abstract
In Winter 2003-2004, the PREMIS working group conducted a survey aimed at gathering information on key aspects of planned and existing preservation repositories for digital materials. Survey questions touched on a variety of areas, such as mission, funding, preservation strategy, and access policies, but with an overarching focus on current practice for managing preservation metadata in digital archiving systems. Survey responses were received from nearly fifty institutions located in thirteen countries, and included libraries, archives, museums, and other institutions. Analysis of the responses suggests that the digital preservation community is beginning to coalesce around several emerging trends in the use and management of preservation metadata, which are enumerated and discussed at the conclusion of the report.

Suggested citation:
OCLC/RLG PREMIS Working Group. 2004. "Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community." Report by the joint OCLC/RLG Working Group Preservation Metadata: Implementation Strategies (PREMIS). Dublin, O.: OCLC Online Computer Library Center, Inc. Available online at: http://www.oclc.org/research/projects/pmwg/surveyreport.pdf (PDF:668K/66pp.)

The following members of the PREMIS Implementation Strategies Subgroup contributed to this report:

George Barnum
Charles Blair
Olaf Brandt
Priscilla Caplan, Chair
Robin Dale
Keith Glavash
Rebecca Guenther
Cathy Hartman
Nancy Hoebelheinrich
Mela Kircher
John Kunze
Brian Lavoie
Vicky McCargar
Evan Owens
Angela Spinazze
Stefan Strathmann
Robin Wendler
Hilde van Wijngaarden
Deborah Woodyard

With staff assistance from Jennifer Childree, FCLA.

TABLE OF CONTENTS

EXECUTIVE SUMMARY 5

Introduction 9

1 Characterization of Respondents 11

2 Mission and Services 13
- 2.1 What is the mission of the preservation repository? Is the repository strictly dedicated to long-term preservation only or does it serve other goals too? 13
- 2.2 Who can deposit materials into your preservation repository? 14
- 2.3 What services does your preservation repository provide? (Check all that apply.) 16
- 2.4 Does your repository manage and track non-digital versions of digital materials as well? 17
- 2.5 How is the preservation repository funded? (Check all that apply.) 18

3 Models and Policies 18
- 3.1 What types of materials are accepted by this preservation repository? 18
- 3.2 How are materials obtained by the preservation repository? (Check all that apply.) 19
- 3.3 What kinds of agreements does the repository have for obtaining materials? (Check all that apply.) 19
- 3.4 Are there formal signed contracts or agreements with customers/depositors? 19
- 3.5 What are the policies or practices of the preservation repository regarding access to the materials? (Check all that apply.) 25
- 3.6 How is your preservation repository informed by the Open Archival Information Systems model (OAIS)? What features of the repository would you say conform to OAIS? What features, if any, do not? 26
- 3.7 In which ways do you find the OAIS reference model useful? Are there ways in which the model is unhelpful? 27

4 Architecture, Storage, and Preservation Processes 28
- 4.1 What is the relationship between access and preservation copies of materials stored in your preservation repository? (Check all that apply.) 28
- 4.2 (Answer only if both access and preservation copies are stored) Do access copies get preservation treatment (e.g. migration)? Is the relationship between access and preservation copies maintained by the repository? If so, how? 29
- 4.3 How are metadata and materials stored within the preservation repository? For example, one repository might zip metadata together with content files and store the zip file as a single entity. Another repository might store metadata in relational database tables and content files as individual entities in a file system. A third repository might use multiple models for different types of materials. Please describe all of the models that apply. 33
- 4.4 What preservation strategies are your preservation repository implementing now? (Check all that apply.) Why did you chose these particular strategies? . 35

4.5 What preservation strategies are your preservation repository planning to implement in the future? (Check all that apply.) Why did you chose these particular strategies? 39
4.6 What type of applications software is in use in your preservation repository? (Check all that apply.) 41

5 Metadata 44
5.1 What categories of metadata are (or will be) stored by and used by your preservation repository? (Check all that apply.) 44
5.2 If you are using or planning to use metadata elements from one or more published scheme, which schemes are you using? (Check all that apply.) 45
5.3 Does your repository record information about these types of entities? (Check all that apply.) 47
5.4 How is metadata obtained (or expected to be obtained) by the preservation repository? For example, is it submitted by depositors, extracted automatically by the repository's computer programs, other? If different methods are used for different sets of metadata, please note all of them. 49
5.5 Do you require contributors to your preservation repository to provide metadata with their contributions? If so, what are your requirements? Please provide sample documentation if possible. 50
5.6 How is metadata stored and updated in your preservation repository? If multiple methods are used, please explain. 51
5.7 Some preservation repositories create or plan to create normalized or migrated versions of digital materials. If this applies to your repository, how will metadata for the original and new versions be handled? What information (if any) will be recorded only for the version created by the repository? 51
5.8 What metadata do you feel is most important to record for preservation purposes? 52

6 DISCUSSION 53

7 REFERENCES 55

APPENDIX A: Alphabetical List of Survey Respondents 56

APPENDIX B: Survey Instrument - PREMIS Implementation Survey 57

APPENDIX C: Follow-up Interview Questions 64

EXECUTIVE SUMMARY

In June 2003, OCLC and RLG convened a Working Group, Preservation Metadata: Implementation Strategies (PREMIS), to focus on the practical aspects of implementing preservation metadata in digital preservation systems. The group has an international membership drawn from library, academic, museum, archive, government, and commercial sectors. One of the tasks in the PREMIS charge is to examine and evaluate alternative strategies for the encoding, storage, and management of preservation metadata within a digital preservation system.

To this end, in November 2003, the group distributed a survey about practices in digital preservation archiving. The survey included questions about business plans, policies, architecture and preservation strategies as well as metadata practices. Copies were sent directly to approximately seventy organizations known to be active or interested in digital preservation. The survey was also made available through several discussion lists. In February 2004, the group made a second distribution, sending thirteen more copies to museums, museum organizations and art institutes, and announcing the survey on lists targeting the museum community.

A total of forty-eight complete survey responses were received. Sixteen of the respondents were later selected for follow-up interviews to gather more information about some practices. Responses came from twenty-eight libraries, seven archives, three museums, and eleven other types of institutions. (A complete list of survey respondents is provided in Appendix A.) Responses were received from thirteen different countries; 46% were from the U.S. Just under half of the organizations had at least some part of their preservation repository in production, while 70% reported being in some stage of planning or development. However, of these, only eleven institutions appeared to have realized an active preservation strategy (migration, emulation, normalization) in production.

Significant results include the following. (Numbers in parentheses refer to sections of this report containing these conclusions.)

> The cultural heritage community has very little experience with digital preservation. We do not have enough experience to indicate whether the metadata these systems record, or plan to record, are adequate for the purpose. (1)
>
> Most repositories serve the two goals of preservation and access. Less than a fifth could be called “dark archives.” (2)
>
> National libraries, archives, university libraries, state libraries, and art museums all seem to have clearly defined constituencies in terms of who can deposit and make use of archived materials, as well as coherent sets of materials of concern. (2)

90% of respondents funded their repositories from their operational budget, while two-thirds used internal or external grant funds, in addition to, or instead of, operations funds. (2)

Differences between libraries and archives in terms of materials accepted are significant and reflect the difference in mission. All archives accepted electronic records and the majority accepted datasets and audio/video. Libraries showed less support for datasets and audio/video but more support for locally digitized materials and web resources. (2)

More than half of respondents had, or planned to have, formal signed agreements with depositors. Most of the agreements contain language that attempts to describe the uses to which repository content can be put, rather than describing in detail the mechanisms of preservation. (3)

Most respondents claim to have been informed by the Open Archival Information System (OAIS) framework, and most say that they at least partly conform to the model. Definitions of OAIS compliance vary, and there is strong demand for supplementary materials including reports and manuals supporting implementation. (3)

All respondents offered “secure storage” as a service; 92% offered, or planned to offer, preservation treatments, defined as normalization, migration, emulation, or other strategies designed to ensure long-term usability. (2)

The majority of institutions chose more than one strategy for preservation. Most (85%) are offering bit-level preservation. Beyond that, restrictions on submissions, normalization, migration and migration-on-demand are the four most popular strategies, in that order. According to the respondent’s future plans, the four most popular strategies, in order, will be migration, normalization, restrictions on submission, and migration-on-demand. Emulation is being used now by only 10% of respondents, but that doubles when future plans are considered. (3)

Most repositories are using some combination of commercially available, open source, and locally developed software. By far the majority are using a combination of software applications. Seventy different commercial and open source software products were specifically named. (4)

Most respondents are recording a wide range of types of metadata; more than half are recording elements of rights, provenance, technical, administrative, descriptive, and structural metadata. (5)

For non-descriptive metadata, METS (Metadata Encoding and Transmission Standard) was by far the most commonly used metadata scheme: 64% of libraries, 42% of archives, and 35% of other institutions used, or planned to use, METS. Z39.87 (Technical metadata for digital still images) was widely used by libraries but not others. (5)

Basing a local metadata scheme on other existing schemes is common, as is using more than one scheme in the repository. (5)

Half or more of respondents record metadata about collections, logical objects (such as books or photographs), files, bitstreams, and metadata. Repositories accept metadata supplied by depositors (75%), extracted programmatically from submitted objects (75%), and supplied by staff (63%). The types of metadata supplied by submitters were predominantly descriptive (64%) with a much smaller portion administrative or technical (22%). (5)

Relationships between files stored in the repository (e.g. access and preservation copies, versions, derivatives, etc.) are maintained in many different ways: through metadata in database tables, metadata in XML schemas, directory structures, file naming conventions, and identifiers. Repositories vary widely in the types of relationships they record. (4)

There appear to be four commonly used models for storing metadata (with the exception of structural metadata) within the repository system. An emerging best practice appears to be to store content data objects in a filesystem or content management system, and store metadata redundantly in a database and with the data objects. Metadata in the database are used by the repository system for operations, while metadata stored with the objects make the objects self-identifying for preservation purposes. In other models, metadata are stored in either a relational database, an XML database, or in a relational database and partially replicated with the objects. (4)

Of the repositories planning to create normalized, or migrated versions of submitted materials, nearly all (95%) indicated that multiple versions (originals and subsequent manifestations) would be maintained in the repository, and that metadata would be created and maintained for all versions. (4)

Trends and conclusions

The following appear to be trends in practice that may ultimately emerge as best practices:

Store metadata redundantly in an XML or relational database and with the content data objects. Metadata stored in a database allows fast access for use and flexible reporting, while storing them with the object makes the object self-defining outside the context of the preservation repository.

Use the METS format for structural metadata and as a container for descriptive and administrative metadata; use Z39.87/MIX for technical metadata for still images.

Use the OAIS model as a framework and starting point for designing the preservation repository, but retain the flexibility to add functions and services that go beyond the model.

Maintain multiple versions (originals and at least some normalized or migrated versions) in the repository, and store complete metadata for all versions. Retention of the original reduces risk in case better preservation treatments become available in the future.

Chose multiple strategies for digital preservation. There are good reasons to have more than one approach in a developing field.

Additional conclusions are listed in Section VI.

Introduction

Background

In June 2003, OCLC and RLG convened a Working Group, Preservation Metadata: Implementation Strategies (PREMIS), to focus on the practical aspects of implementing preservation metadata in digital preservation systems. The Working Group has an international membership drawn from library, academic, museum, archives, government, and commercial sectors. In addition, it has an international Advisory Committee that provides expertise in periodically reviewing progress and providing feedback.

The PREMIS objectives are to:

- define an implementable set of "core" preservation metadata elements, with broad applicability within the digital preservation community;
- draft a data dictionary to support the core preservation metadata element set;
- examine and evaluate alternative strategies for the encoding, storage, and management of preservation metadata within a digital preservation system, as well as for the exchange of preservation metadata among systems;
- pilot programs for testing the group's recommendations and best practices in a variety of systems settings;
- explore opportunities for the cooperative creation and sharing of preservation metadata.

To accomplish these objectives, the Working Group divided into two subgroups, the Core Elements Subgroup and the Implementation Strategies Subgroup. The Core Elements Subgroup is charged with defining the core metadata element set and drafting the data dictionary. The Implementation Strategies Subgroup is responsible for the examination and evaluation of strategies to manage and exchange preservation metadata, and for piloting the Core Elements Subgroup's recommendations.

Several reports have described the state of digital preservation in the arts and sciences. The Library of Congress' *Plan for the National Digital Information Infrastructure and Preservation Program* contains a wealth of background information on both national and international efforts (NDIIPP 2002). A survey of Digital Library Federation members in February 2002 showed that few respondents had formal digital preservation policies in place, but several had begun preservation-related activities (Flecker 2002). "It's About Time: Research Challenges in Digital Archiving and Long-term Preservation" sponsored by the NSF and Library of Congress, examined the state of digital archiving to ascertain the most pressing research needs of the community (NSF 2002). "Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation" made similar observations a year later (NSF 2003). "Digital Preservation and Permanent Access to Scientific Information: The State of the Practice" by CENDI and ICSTI focused on the state of operational digital preservation systems in science and technology specifically (CENDI 2004). For that report, more than fifty

archiving systems or projects were identified, twenty-one of which were selected for in-depth interviews.

While the CENDI report, in particular, provides a wealth of information about various stakeholders, important document formats, and approaches to preservation, no report to our knowledge provides a detailed picture of how preservation repository systems are actually being implemented. To this end, in November 2003, the Implementation Strategies Subgroup distributed a survey about practices in digital preservation archiving. The survey instrument is included as Appendix B.

To develop the survey, the group first drafted a list of topics to explore, such as:

- What are the missions of preservation archives in terms of materials they will archive?
- How are metadata for migrated objects being treated?
- Can we identify general models for storage of content objects and metadata?

Although the charge focused on metadata, the group felt that the survey provided an opportunity to explore the state of the art in digital preservation generally, and the final list of topics included areas such as policies, governance and funding, system architecture, and preservation strategies. A set of survey questions was then drafted to elicit information about the topics on the list. Questions were closed-ended whenever possible (that is, they offered a finite set of answers for respondents to chose from) although many questions also included a prompt to explain, give examples, or otherwise elaborate on the answer. A typical hybrid question looked like this:

> 5.6. How are metadata stored and updated in your preservation repository? If multiple methods are used, please explain.
>
> __ in a relational database
> __ in an XML database
> __ in an object-oriented database
> __ in a proprietary database or format
> __ in flat files
> __ bundled with related content files

The instrument was pretested on nine PREMIS institutions, tweaked and finalized. Finally, in November 2003, copies were sent by email directly to approximately seventy organizations thought to be active in, or interested in, digital preservation. The survey was also made available on the PREMIS web site, and general invitations to respond were posted to DIGLIB, ERECS-L, METS, OAIS-IMPLEMENTERS, and DPC-DISCUSSION mailing lists. In addition, the Digital Preservation Coalition redistributed the survey to institutions in the U.K. In January 2004, follow-up emails were sent to contacts at any of the original seventy organizations that had not already responded. In February, the group made a second distribution, sending thirteen more copies to museums, museum organizations, and art institutes, and announcing the survey on lists targeting the museum community.

By the end of March 2004, forty-eight survey responses were received from institutions developing or planning to develop a digital preservation repository. After reviewing

responses, the subgroup identified a number of areas where more detail would be helpful. A list of follow-up questions was drafted, and sixteen respondents were contacted for telephone interviews.

Answers to the closed-ended questions were tabulated. Subgroup members then divided up responsibility for analyzing responses to one or more survey questions and any related interview questions, and for drafting this report. The results described in this report are based on forty-eight survey responses, sixteen telephone interviews, and a small number of ad hoc emails exchanged to clarify specific points.

1 Characterization of Respondents

Survey responses came from national libraries, state libraries, university and research libraries and consortia, archives, museums, and a few organizations that fall outside of these categories. Categorizing the forty-eight responses is not entirely straightforward because some institutions had dual functions. For example, one institution is both a national library and national archives. Another is both a national library and a university library. In these cases, the institution was counted in both categories, so the total adds up to more than forty-eight.

In other cases the response came from a subunit of a larger institution. In this case the institution was classified at the lowest level, so that, for example, an art institute associated with a university was counted as a museum. Finally, one institution sent two survey responses, one describing their current preservation repository and one describing their future plans, because the future plans were quite different. We counted this as two responses, from two institutions. All other respondents described current implementations and future plans in the same response.

Category	Subcategory	Total	Percentage
Libraries		28	58%
	National Libraries	11	23%
	Academic Libraries/Consortia	15	31%
	State Libraries	3	6%
Archives		7	15%
	National Archives	6	13%
	Institutional Archives	1	2%
Art Museums	Art Museums	3	6%
Other		11	23%
	U.K. National Center	4	8%
	U.S. National Center	1	2%
	U.S. Government Agency	1	2%
	Newspaper	1	2%
	Not-for-profit organizations	4	8%

For the purposes of summarizing differences between different types of institutions we used a simpler breakdown as follows:

Institution	Count	Percentage
Libraries	28	58%
Archives	7	15%
Other	14	29%

The majority of responses came from libraries, archives, and research councils or other agencies associated with higher education. Sixteen responses came from members of the PREMIS Working Group or Advisory Committee, representing just over half of the membership, which is not surprising given the constitution of the group. Overall there were twenty-five replies in response to the original mailing to seventy institutions, a 35% response rate, and no replies in response to the February mailing. The remaining twenty-three responses presumably came from indirect solicitations. Despite the separate solicitation to museums, only three art museums, and no natural history museums, sent replies. Several museums that were contacted directly to participate declined to do so because they were not yet actively engaged with planning or implementing a preservation repository.

Although several institutions known to be developing digital preservation repository systems did not respond, we believe the replies received were reasonably representative of the state of the art in the winter of 2003/2004.

Responses came from thirteen different countries. Most came from the U.S., followed by the U.K. A breakdown of responses by country is provided below.

Country	Count	Percentage
Australia	2	4%
Austria	2	4%
Canada	2	4%
France	1	2%
Finland	1	2%
Germany	2	4%
Netherlands	1	2%
New Zealand	2	4%
Portugal	1	2%
Sweden	1	2%
Switzerland	1	2%
United Kingdom	9	19%
United States	22	46%
TOTAL	48	100%

Nearly half of the organizations responding had at least some part of their preservation repository in production, while 40% were only in the planning stage. Numbers add up to more than 100% because several organizations reported being in multiple stages at

once. For those not yet in production, target dates for production implementation ranged from 2004 to 2007.

Preservation Repository State	Count	Percentage
Planning and Organizational stage	18	38%
Development (alpha, beta)	16	33%
Production	22	46%

For the purpose of categorizing other responses, we assigned each response to a single category, "planning/organization" or "development/production," based on an overall assessment of the repository. That breakdown is as follows:

Preservation Repository State	Count	Percentage
Planning/organization	15	31%
Development/production	33	69%

These results indicate that there is very little experience with digital preservation. Twenty-two respondents claimed to have a preservation repository in some stage of production (as opposed to planning, development, or alpha/beta testing). However, only half of them appeared to have implemented an active preservation strategy such as normalization, format migration, migration on demand, or emulation. These included four national libraries/national archives, and six institutions categorized as "other." None was an academic library.

This finding must color all other results, including those pertaining to metadata. Whatever practices were reported in the survey, apart from these eleven institutions, the results reflect repositories not yet in production, or not yet implementing active preservation strategies. In the remainder of this report, results and analysis are arranged under numbered survey questions, in the order of the survey.

2 Mission and Services

2.1 What is the mission of the preservation repository? Is the repository strictly dedicated to long-term preservation only or does it serve other goals too?

Most repositories serve the two goals of preservation and access. Twenty-one surveys specified preservation and access explicitly in their replies, while another five used the word "dissemination" instead of, or in addition to, access. Since forty-one respondents indicated that they offer online real-time access to service or archival copies, we can assume that many of the repositories that did not mention access as a goal had access as a function. Only four repositories indicated they serve the goal of preservation only, with access supported by a different system.

Other goals commonly mentioned were storage (two), discovery (three), data management (three), and acquisitions/collection (three). At least five responses described a set of goals and functions that could be summarized as the objectives of an academic institutional repository. These include the self-archiving of institutional

research such as article preprints and postprints, theses, and dissertations; management of digital collections; preservation of digital materials; housing of teaching materials; and electronic publishing of journals and books. (Cervone 2004)

Some goals mentioned were clearly goals of the larger organization rather than of the preservation repository itself. These include interpretation of data, promotion of the use of data, recording and monitoring external publications, and encouraging publisher contributions. A national archives included among its goals the guarantee of authenticity; another archives listed the goal of short-term storage for official records until they could be disposed of. A newspaper repository had the goals of research support for journalists, and revenue via export of content to database vendors. Two repositories listed repurposing and reuse of content.

2.2 Who can deposit materials into your preservation repository?

Depositor	Libraries	Archives	Other	Total	Percent
General public	5	0	1	6	13%
Research community	12	0	5	17	35%
My institution	17	3	7	27	56%
Other companies	13	0	6	20	42%
Subscribers	1	0	1	2	4%
Other	7	4	4	15	31%

The answers here were fairly consistent within the various categories of respondent. Viewed together with answers to question 3.1 (*What types of materials are accepted by this preservation repository?*) a picture emerges of different types of institutions developing preservation repositories around the materials, needs, and organizational/legal environment of their constituencies.

2.2.1 National libraries

National libraries generally saw themselves as the primary depositors to their own repositories. A typical answer was that library itself was the only depositor, but that the library could obtain materials by purchase, donation, deposit, agreement, harvest, or in-house creation. One national library was in the early development stage and had not yet formulated its policies. Of the remaining nine, all reported archiving materials from national deposit programs. Five had required legal deposit programs for digital materials, three had voluntary deposit programs or agreements with publishers, and one was drafting new deposit laws. Most (but not all) legal deposit programs were limited to offline materials (e.g. CD ROM); one national library reported a legal deposit program for offline materials and voluntary agreements with publishers for online materials.

Most (6) national libraries mentioned archiving materials digitized in-house; one also accepted materials digitized by external cultural heritage institutions. Two had arrangements with universities for archiving electronic theses and dissertations (ETDs) and other materials while a third were planning this for the future. Two had formal agreements to share the preservation repository with partner institutions. At least two were harvesting web sites by agreement with the publisher.

2.2.2 Archives

Of the seven institutions categorized as archives, six were national archives and one was the archives of a federal institution. National archives allow the same parties to deposit digital materials as deposit paper records: generally government agencies and some private donors. The institution's archives accepted deposits only from the institution. In all cases the repositories intended to preserve records and related documentary materials.

2.2.3 University and Consortia Libraries

University libraries tend to accept materials only from their own institution, but three respondents mentioned that they also accepted materials from their partners in digitization projects. Of the fourteen university libraries responding, eleven appeared to be planning to archive "institutional repository" type content - research and scholarship materials produced within the university system, including by the library itself. In most cases the library planned to exercise some selection criteria on what would be accepted by the repository.

Two respondents indicated they would take anything submitted by an organizational entity of the university or consortium, as long as the submitter had the right to authorize deposit and preservation actions. Interestingly, both of these repositories charged or planned to charge for use, raising the question whether these respondents were using cost instead of selection as a filtering mechanism. A number of repositories also accepted materials from outside the university. Two intended to select materials of long-term research value from any source, one said they would archive web sites pertaining to their region, and one would take materials from any institution in the educational system. One university library had an agreement with the state to archive federal and state publications.

2.2.4 State libraries

Two American state library agencies and the library of one Australian state responded. The two U.S. libraries accepted materials from the library itself and from government agencies. Both of them intended to archive official state government publications, while one also collected federal publications. The Australian library was a regional/jurisdictional repository for documents, including digital documents, from government, commercial and community publishers within the state.

2.2.5 Museums

Three art museums responded to the survey. Of these, two were explicitly interested in media art. One accepted submissions only from the museum itself, including digital content generated in-house and digital art works purchased or donated by artists and collectors. The other appeared to have a similar policy, except that it also required candidates for a media arts prize to agree to deposit their art in the repository. The third museum archived only digital surrogates of physical objects owned by the museum.

2.2.6 Other

The remaining responses included four U.K. national service centers with missions to support and promote certain types of research. All of these had similar policies, allowing deposits from the research community as well as the organization itself or selected

external organizations. All of them sought publications or datasets within their area of scope (e.g. social science datasets, atmospheric data sets).

The most common use of the "Other" category was to note government branches and agencies. The "Other" category was also used by several academic institutions including, or considering the inclusion of, research materials of interest to their own communities. Their comments included:

> Other: when they have produced a digital resource of significant potential value to our users. The 1881 Census dataset for the U.K. is one example.
>
> Our own institution, institutions that are part of our Educational System, and others with whom we are working on projects. The research community as a whole is under evaluation.

Several respondents clarified that their own institution deposited content obtained from external sources, or reviewed external submissions. Examples are:

> Digital material can be purchased, donated, acquired by legal deposit, harvested from the Internet, or created by the [library]. At this stage only the [library] can deposit materials received by any of these methods directly into the preservation repository.
>
> All state agencies, boards and commissions. However, all deposits will be screened during the in-take procedure.

2.3 What services does your preservation repository provide? (Check all that apply.)

Service	Count	Percentage
Search and Discovery	38	79%
Online, real-time access to service copies	39	81%
Online, real-time access to archival copies	28	58%
Secure storage	48	100%
Data Management	44	92%
Storage and/or management of non-digital	14	29%
Preservation	44	92%
Formal distribution	35	73%
Reporting	34	71%
Billing	11	23%
Other	8	17%

All but seven respondents checked either "online, real-time access to service copies" or "online, real-time access to archival copies" or both. This corresponds perfectly with answers to question 4.1 below, where exactly seven respondents indicated that no access copies were stored or generated. This indicates about 15% of respondents do not offer any routine online access, and might be called "dark archives."

Most comments were clarifications of categories that were checked, and included the following:

> Online, real-time access to service copies will be restricted to some of our holdings (i.e. not extremely large datasets without subsetting; not extremely large image/audio/video files).
>
> Online, real-time access to archival copies will be restricted to staff and subject to various limitations.
>
> It will aid discovery, but not provide detailed search facilities.
>
> [The archive] currently creates an archival ("master" or "preservation") copy consisting of a master and backup master copy, on preservation media. On request, an access or reference copy is made from either the master or backup master copy.
>
> [Although we checked all of the above,] various aspects may not yet be implemented (e.g. preservation treatments), may not apply to all items (e.g. management of non-digital versions, billing), or may be available only through library staff (e.g. access to archival copies). "Other" aspects for implementation include rights management.
>
> We have not yet decided whether online access to digital records will be provided.
>
> This year preservation functionality will be added, which will contain the management of technical file format information and keeping track of the consequences of technology changes.

Some respondents noted other services offered:

> [We offer] a Submission Builder to allow users to create a METS encoded submission information package. We also offer a harvester for Web documents.
>
> Support for bulk operations on the archived materials including registration, loading, unloading, and retrieval.
>
> To ensure the integrity of data, we plan to give a possibility to compare checksums between the live repository and the long-term preservation archive.

2.4 Does your repository manage and track non-digital versions of digital materials as well?

	Count	Percentage
Yes	15	31%
No	31	65%
Not yet sure	1	2%
No answer	1	2%

The relatively high proportion of repositories tracking non-digital versions was surprising. A third of these (five) were archives, some of which noted they were obligated to take records in any format. Several respondents reported that non-digital versions were managed by other units within the same organization, or were handled by systems other than the repository system. Five of these noted that they provide links between the records for digital and non-digital versions. In two cases this was considered tracking the non-digital versions and in three cases it was not.

One of the "no" answers elaborated:

> Non-digital versions of digital material have their own management stream. Whilst digital and analogue objects may share descriptive metadata they have significantly different properties and characteristics that mean their management processes must be different. Relationships between digital and non-digital versions of the same object are managed through the metadata.

2.5 How is the preservation repository funded? (Check all that apply.)

Type of Funding	Libraries	Archives	Other	Total	Percentage
Grant funded externally	14	1	6	21	44%
Grant funded internally	8	1	1	10	21%
Fee for service	4	1	2	7	15%
Operational budget	24	7	12	43	90%
Other	4			4	8%

By far the majority of preservation repository projects and services are funded entirely, or in part, from the organization's operating budget. Several national archives noted that funding came from the government and was considered operating budget. Twenty-two institutions (45%) have more than one funding source. For all of them, the operational budget was one of the sources. Only four institutions indicated that their repositories were solely supported by grants. Nine respondents indicated modes of funding would change in the future. Two of these said they were looking for external grant funding, and three said they would or might move to a fee-for-service model. The other four did not specify the nature of the change.

3 Models and Policies

3.1 What types of materials are accepted by this preservation repository?

Materials	Library	Archive	Other	Total	Percentage
Electronic records and/or publications	22	7	11	39	81%
Datasets	8	5	4	17	35%
Digitized material	12	1	12	25	52%
Audio/Video	12	4	5	21	44%
Web resources	10	2	7	19	40%

Differences between libraries and archives were striking, reflecting the difference in mission. Respondents for archives generally noted that the archive was required to take certain records regardless of the formats in which they were produced. They all accepted electronic records, and the majority accepted datasets (71%) and audio/video (57%). Only one accepted locally digitized material, presumably because retrospective digitization of legacy records has not been a priority for archives. Libraries, in contrast, showed less support for datasets (28%) and audio/video (42%) but more support for locally digitized material (42%) and web resources (35%).

3.2 How are materials obtained by the preservation repository? (Check all that apply.)

Method of Obtaining	Count	Percentage
Harvested by repository	23	48%
Submitted to repository	46	96%

Nearly all respondents obtained materials by submission. Most used submission alone (52%), or both submission and harvesting (45%). Only two institutions, (4%) used only harvesting. National libraries most heavily used harvesting. Many of these have agreements with publishers to harvest ejournals or selected web sites. Submission was strongly preferred by archives. Two institutions commented that they did not currently harvest but that they were planning to add this function.

3.3 What kinds of agreements does the repository have for obtaining materials? (Check all that apply.)

Agreements	Count	Percentage
Customers chose what to deposit	21	44%
Governmental deposit agreement	16	33%
Institutional archiving agreement	22	49%
Other legal mandate	10	21%
Other	12	25%

Four of the respondents checking “Other” described negotiated agreements with resource providers including contracts with publishers. Other “others” included:

- Web site owners asked for permission
- Memorandum of Understanding
- Title 44 U.S. Code

3.4 Are there formal signed contracts or agreements with customers/depositors?

	Libraries	Archives	Other	Total	Percent
Yes	10	5	10	25	52%
No	14	0	2	16	33%
N/A	1	2	2	5	10%

More than half of the respondents had, or planned to have, formal signed agreements with depositors. All but two archives reported having formal agreements, both exceptions claiming this to be not applicable, possibly because law requires submissions. Institutions in the "Other" category also highly favored signed agreements (71%), possibly because national centers and private not-for-profits deal with outside customers. An almost even split among libraries (42% yes, 50% no) reflects the difference between libraries dealing with external publishers and agencies, and libraries dealing mainly with their own content.

Some "yes" answers were qualified:

> In the case of video or a guest speaker or donated artwork, but not for content generated entirely in-house.
>
> In some cases formal agreements are entered into but they are not essential in every case. For digital material the same types of deposit agreements will exist as are currently used for analogue material.
>
> [The repository] does not have a formal, legal agreement with communities that submit material, but does have written policies on institutional rights and responsibilities for both submitters and repository managers, and requires submitters to click through a license agreement which spells out some of these rights and responsibilities for each submission.

Many of the institutions that used agreements contributed samples, which provided a great deal of useful information. Analysis of the samples gave us some specific information about how different self-described preservation repositories expressed:

- stakeholders associated with the repository's function and services;
- rights that were granted both to the submitter and to the repository itself;
- preservation actions the repository planned to take, and the permissions needed to perform those actions;
- subsequent warranties made based on the preservation actions;
- restrictions imposed upon a user of the repository's content;
- conditions to which a user of the repository's content must agree.

From the examples of terms, it appears that most of the agreements contain language that attempts to describe or constrain the uses to which the repository content can be put, rather than describing in much detail the mechanisms of preservation. When the agreements were concerned with the topic, the right was often couched in terms of the institution's right to ***preserve*** the content it receives.

Many of the agreements also explicitly express the institution's right to provide access to the repository's content although this is not nearly as common, especially since some of the repositories consider themselves "dark," with no intent to provide access to those not identified as coming directly and authoritatively from the content submitter. Language pertinent to "access" is not addressed here except when it is specifically described in relation to preservation activities.

Conditions and restrictions upon use described in the agreements seem to be important in relation to the institution's efforts to protect it against copyright violations, and so will be discussed in that context.

3.4.1 Rights “to preserve”

For most institutions, the strategy for describing the rights granted for preservation activities seemed to use broad terms that allow for changes in preservation methods and techniques over time as technical solutions and social expectations change.

> To preserve and make accessible in a variety of formats and media; archive, distribute and use.
>
> Right to convert from one file format to another (presumably for converting to Unicode).
>
> Take the necessary preservation actions to keep publication accessible as hardware and software changes over time.
>
> Provide Archive Services at full, bit preservation or local service levels as defined and documented over time.
>
> Non-exclusive right to reproduce, translate and/or distribute data (including abstract) worldwide in print & electronic format and in any medium including, but not limited to audio or video.
>
> Nonexclusive, nontransferable, and nonassignable right to make use of documented services of the Archive (which does not include access).

3.4.2 Repository actions

Often, when more precise terms were included in the agreements, they were more related to actions the repository could perform than to specific Rights. Examples of those terms follow:

> Electronically store, translate, copy or re-arrange to ensure future preservation and accessibility.
>
> Make necessary copies of the data for purpose of preserving.
>
> Ensure adequate physical custody, validation, dissemination & review/purging of data.
>
> May copy deteriorated or damaged documents after they have been copied in a form that retains all the information in the original documents.
>
> Store and manage titles according to published preservation policies.
>
> Use, copy, display and prepare derivative works from data and from metadata about the data to provide services at proscribed levels of service.

Translate, without changing the content the data to any medium or format for the purpose of preservation.

May keep more than one copy of data for purposes of security, back-up and preservation.

Repository will not make any alteration, other than as allowed by agreement to the submission.

May dispose of records having no long-term value after specified retention period, or transfer / keep records in Archive for long-term preservation.

To store, translate, copy or re-format the data in any way to ensure its future preservation and accessibility.

Use the best standards and procedures for the storage, manipulation and access of digital materials as it evolves over time. Employ appropriate technical solutions including the latest technologies in cooperation with the depositor to ensure continued availability of the data.

In follow-up interviews, respondents were asked more specific questions aimed at finding out whether repositories felt it important to have rights granted to them which expressly allowed either migration or preservation activities. Most of those institutions contemplating migration as a preservation strategy seemed to consider the agreements they had developed either adequately addressed that particular strategy without naming it as a method, or were in the process of planning that such coverage be included in the agreements. In contrast, very few institutions were actually contemplating the use of emulation, and if so, they felt that the language in their existing agreements would cover its use. The only exception to the rule was an art museum, which collected variable media art requiring specific permissions from the rights holder to perform emulation as well as explicit instructions for performing the emulation. Those explicit instructions were considered to be part of the license agreement as terms and conditions necessary to initially display the artwork, but also for longer-term preservation, i.e., re-creation of the artwork at a later time. Although we have no other samples of agreements from museums, this exception may well prove the rule for museums as contrasted to digital libraries or digital archives.

3.4.3 Warranties regarding preservation responsibilities

In some agreements, institutions used warranty clauses to further clarify the extent of their preservation responsibilities with their content submitters. Sometimes the warranties are made by the submitter, and sometimes by the repository itself. Most frequently, the warranties made by the submitter have more to do with the assertion of copyright holder than anything to do with preservation activities. When language does refer to preservation activities, it often deals with the condition of the content. In these cases, it is not clear whether the institution is trying to protect itself against what might happen as a result of preservation activities, or whether they are merely addressing the condition of the content as received from the submitter. Some examples of warranties made by the preservation repository follow:

> No obligations to reproduce, transmit, broadcast, or display in same formats or resolutions as received.
>
> Archive has no liability for physical materials deposited or for loss of data or information in the operation of the Archive.
>
> Data collected not warranted to be suitable for use by recipient.
>
> Archive maintains integrity and long-term accessibility of Archive and Subscriber's Content Objects within [the archive].
>
> [repository] will provide for the permanent storage and maintenance of data in a form that will provide security to data integrity and usability; will maintain the content of the data, not the format or functionality of the content.

3.4.4 Protection against copyright violation

Interviewees were also asked about the institution's practices for protecting themselves and their submitters against possible copyright violations. When asked if the repository was relying on submitters holding copyright in order to grant intellectual property rights to the repository against possible violation of applicable copyright laws, all but one answered positively. The other institution did not reject this position, but rather noted that the matter was still being explored.

Another follow-up question asked whether internally digitized digital materials were treated differently from external materials in terms of documentation of the copyright holder (since it was presumed that copyright holder would be the institution itself, or a parent organization). In these instances, institutional practice was much more disparate, when the situation applied (for six out of fourteen, the situation was *not* applicable). Three institutions had not yet developed a consistent practice. Other institutions either relied upon explanatory documents, which accompanied the digital object, embedded the pertinent information in the digital object itself, or put it in a metadata record that displayed with the object.

There appears to be an implicit link between the declaration, or warranty about who owns copyright, and the use conditions or the restrictions upon the user of the digital content. That is, the use conditions and user restrictions often help clarify the nature of the copyright agreement, and should be considered a useful method of documentation in addition to any warranties within a license or contract, and any rights metadata.

Some warranties made in formal agreements regarding copyright were:

> Depositor is copyright holder or authorized by; data collection doesn't violate copyright laws.
>
> Repository is under no obligation to take legal action on behalf of rights holders if IP rights or other rights breached.
>
> Licensee must be owner of copyright or “duly authorized” representative of owning institution.

The archive warrants the depositor retains copyright.

Original owner retains copyright and Intellectual Property rights of the data, software, information, or other documentation.

Depositor certifies that any restrictions on use are in conformance with requirements of pertinent law.

Depositor has authority to grant permission to archive on behalf of all contributors to data being archived; Contributors to documents being archived have been notified of deposit and agree.

Subscriber maintains ownership in content; Subscriber is responsible for complying with copyright and other laws related to proprietary content.

Depositor has the rights to grant the rights specified w/in the license, and does not infringe upon anyone's copyright, to the best of depositor's knowledge.

Submitter retains ownership rights in deposited data; is responsible for compliance with applicable copyright laws.

Depositor has right to grant permissions in contract as copyright holder, joint copyright holder with authorization to grant license or not copyright holder, but with authorization to grant license.

Agreement of all parties who may have in interest in the data collection has been obtained.

Author warrants that he/she is copyright holder and has received consent to publish elements of the Work to which others have title.

Submission does not infringe upon anyone else's copyright, to the best of the author’s knowledge; author has received unrestricted permission to grant repository the rights required by the license, for those elements not held in copyright by the author.

Agreements contained these examples of restrictions on users and uses:

For non-commercial educational purposes only (teaching, research & private study); must be Authorized User who has agreed to abide by license conditions.

Must protect individual confidentiality of original translators; must not copy data in whole or in part unless it is for user's exclusive use to conduct research, nor allow others to copy data unless they are directly associated or working with user & under same terms as user.

For commercial publication, access limited to physical premises of the Library on single computer with copying and communication functions disabled during time in which publication is commercially viable.

Disclosing, reproducing, distributing or transmitting data in any form except as permitted by terms of license.

No use, which circumvents security measures, implemented by the Archive or interferes with the functioning of any web site or computer system.

Grounds for restricted access specified by depositor, and include criteria, duration, and point from which restriction applies and which files covered by restriction.

Not to breach copyright of copyright holder by selling all or part of the data, or including it in a product subsequently sold.

Author has no right to remuneration for publication.

Examples of Use Conditions from formal agreements

Copyright in original data not transferred.

Required to acknowledge rights holder when publishing anything based on the data; required to state that rights holder is not responsible for the quality of the work users produce; To keep list of all persons to whom access of data has been given; supply a copy of the list along with copies of the written undertakings of those person to Archive Director when asked.

Make processed / derived datasets resulting from project available to Data Centres for licensing / transfer to others.

Must adhere to Privacy Act provisions, if applicable.

Suspension of end-user access to data in case of claim of copyright or other violation; Attach notice of restrictions when making data available to end-users.

Not attempt to identify any individuals whose details appear in the data, where this is relevant.

3.5 What are the policies or practices of the preservation repository regarding access to the materials? (Check all that apply.)

Policies and Practices	Count	Percentage
Open access to all end users	26	54%
Access restricted to a particular community	29	60%
Access after a specified trigger event	22	46%
On site access only	18	38%
Paid access	4	8%
No online access	14	29%
Other	12	25%

Responses to this question did not show great variation between libraries, archives, and other institutions. However, they do indicate at least two things: first, repositories interpret “access” in many different ways, and second, that access policies are rarely simple. The question did not explicitly define access, and most respondents did not clarify what they meant by their answers. However, from the few comments received, it appears that access was variously taken to apply to metadata (that is, access to a catalog or inventory of resources rather than the content itself), to service copies, and to preservation copies.

Interestingly, four-fifths of all respondents checked multiple access policies, and more than half (52%) checked three or more policies. It was almost never the case that a repository simply offered open access to all end users: of the twenty-six respondents that checked this category, twenty-five checked other policies as well. Comments revealed that access policies depended on type of material, type of material in conjunction with category of user, acquired distribution rights, and owners’ specifications. The response of an art museum was typical of the complexity of most comments received:

> Access to in-house staff to everything. Access to researchers and other collaborators on a per-request basis. Access to the general public to ‘access versions’ via the museum web site.

3.6 How is your preservation repository informed by the Open Archival Information Systems model (OAIS)? What features of the repository would you say conform to OAIS? What features, if any, do not?

Two institutions did not answer the question at all; three replied that they could not answer the question because they had not examined the model deeply enough, and one was unaware of OAIS. Most of the remaining responses claimed to be informed by OAIS and said their repositories at least partly conform to the model. Four institutions called themselves fully conformant with OAIS.

One institution claimed that OAIS was a touchstone for their existing archive, but they differed in the conception of some functions. Some functions were implemented outside the archive (e.g. the access function, which in this implementation is unified with access for non-digital materials) and some were implemented in a different way (e.g. descriptive metadata are not tied to archival objects per se). One institution called OAIS very important in the design of their digital archive. The requirements that were given to the developer of the system clearly stated that the design had to comply with OAIS, although they did add functions outside of the OAIS model. One repository partially supports the information model of OAIS: they use the concept of Information Packages, but they have problems with the conceptual bundling of the "data object" with “representation information” as “content information.”

As a follow-up question, respondents who were interviewed by telephone were asked what it meant to be OAIS-compliant. The eight respondents’ answers grouped themselves into three positions: that OAIS-compliance is clear enough in the OAIS paper itself in sections 1.4, 2.2, and 3.1; that compliance means you have equivalent functionality or that you incorporate the main characteristics of the model, such as the six core functional components and the specified responsibilities; that compliance is a

vague concept or that frameworks are nothing you should/could comply to. One institution in the last group commented that the framework is only a starting point and not necessarily something with which to be compliant. Another felt that compliance has little significance in relation to "reference frameworks." Their repository is being built to model the architecture, functional areas, and services described in the OAIS reference model. Given the various ways in which digital repositories have applied OAIS concepts to their infrastructures, it doesn't seem very realistic to expect any "compliance" in real terms.

3.7 In which ways do you find the OAIS reference model useful? Are there ways in which the model is unhelpful?

Three respondents said that OAIS is a good guide to showing the requirements for a digital archive and that it helps in the planning phase. However, the planning of one of these institutions was done before the publication of the OAIS model. One respondent felt that OAIS was a too limited as a concept, but did not elaborate. One respondent had problems with the OAIS vocabulary. Their main concerns were that OAIS terminology encourages the use of less specific terms and concepts over more specific ones, and that it introduces competing definitions for terms already established within the library community, which can lead to confusion. Specific complaints included:

> The congruence between “content data object” and “data object,” and “content information” and “Information object.” Respondents did not see a need for different terms with the same meaning.
>
> The term of “Producer,” defined in OAIS as “The role played by those persons or client systems who provide the information to be preserved.” They suggest “Provider” or “submitter,” because it comes closer to the meaning in practice. Respondents argue that in the library context, a producer is the organization responsible for creating the content of a database, as opposed to the vendor that markets it. The distinction between those who produce information and those who provide it to a preservation repository is important.
>
> The definition of “digital migration” as “the transfer of digital information, while intending to preserve it, within the OAIS. ...” Respondents criticize the word “transfer” (which is not defined in OAIS), which connotes a media migration, but not a transformative migration (a format or forward migration).

One respondent thought that OAIS is a good starting point for community specific interpretations and implementations. They commended the attention to the life cycle of items and the more organizational part of OAIS, like the consideration of standards, users, workflows, and responsible agents. On the other hand, they criticized OAIS for lacking clear definitions of communities of users and not showing how the principles set out in the model might be implemented in those communities. One institution criticized the traditional approach of the OAIS model, which conceptualizes the repository in a “passive” way where materials and associated information from the outside are collected and made accessible. This institution is moving to a model in which the depositing institutions actively take part in the preservation process (e.g. preparing the records, helping to ensure authenticity). One institution noted that OAIS was quite useful for building their digital repository, but they criticized the model for being imprecise in its description of data administration and administration functional entities. They see an

overlap in these fields, which for them make the relations to preservation planning unclear. They wish there was more guidance for specific functional requirements. Additionally they consider the access functional entity as insufficiently developed (with the exception of the DIP construct). They give the example of differently copyrighted materials (some are contracted, some not, some are in the public domain...). What they miss is a best practice discussion for different situations in the digital repository world.

As a follow-up question, interviewees were asked; *do libraries and museums need a different model, or another approach to the OAIS model?*

Most regarded a new model as unnecessary or even undesirable. However, there were some proposals to augment OAIS in some areas (e.g. versioning and relationships between multiple instantiations of an object) and to rewrite the essentials in a way that allows libraries to add annotations. There was also a strong demand for supplementary materials including implementation reports and implementation manuals. Every institution questioned was interested in getting more information about practices and/or best practices of implementations. Some want to have examples for comparability reasons, for learning from different contexts, and to make clear what complying with OAIS means.

There was some feeling that different local needs will lead to different models in implementation practice. One library thought that some work has to be done to flesh out the model in different implementations in different contexts. Another library suspected that different local demands, "pressures," internal structures and roles are spawning different models in practice. An archive predicted an integration of existing standards for archives and OAIS principles. A museum/archive did not see a different model superceding OAIS in the museum sphere, but thought that new models will enhance OAIS by going further and supplying more details. The idea of a lifecycle for a model was brought up. OAIS might be revisited and revised after some experience; perhaps at some point a more meaningful model will replace it.

4 Architecture, Storage, and Preservation Processes

4.1 What is the relationship between access and preservation copies of materials stored in your preservation repository? (Check all that apply.)

Relationship	Count	Percentage
Access and preservation are served from the same copy	23	48%
Access copies are generated "on the fly"	19	40%
Access and preservation copies are stored in the repository	23	48%
Access and preservation copies are stored, but not in the same repository	18	38%
No access copies are stored or generated	7	15%

Two institutions had not yet determined the relationship between access and preservation copies. Twenty institutions selected a single relationship between access and preservation copies (see below for a breakdown). Twenty-six institutions described

multiple relationships between access and preservation copies, using fourteen different combinations.

Institutions checking only one relationship	Count
Access and preservation are served from the same copy	6
Access copies are generated “on the fly”	4
Access and preservation copies are stored in the repository	3
Access and preservation copies are stored, but not in the same repository	6
No access copies are stored or generated	1

The comments of some institutions indicated confusion about the terms used in the survey question. For example, some implied that “on-the-fly” meant automatic conversion at ingest, while for others, it meant conversion when a request for dissemination was received. “Dark archive” is another term that was not applied consistently. One institution did not check “dark archive” but stated that there was no external access to the repository. Since repository staff need to be able to access stored content, the term “dark” can be seen as misleading.

Many institutions did not explain their reason for selecting multiple options. Because of the lack of information, it is not possible to determine the import of some of the combinations. It could be that some options were meant as additional detail about the first relationship. That is, if an institution checked that “access and preservation are served from the same copy,” then the statement that access copies are generated "on the fly" could be taken as additional detail about the use of that copy. Two institutions checked all choices, indicating (we assume) that their systems were able to support all variants of the above scenarios. In any case, it is clear that there were at least two interpretations of “access and preservation are served from the same copy.” To some this meant there was only one copy in the repository, while to others it meant that the access and preservation copies are two separate but identical digital entities.

4.2 (Answer only if both access and preservation copies are stored) Do access copies get preservation treatment (e.g. migration)? Is the relationship between access and preservation copies maintained by the repository? If so, how?

Half of the respondents answered this question. Of these, roughly a third (36%) planned preservation treatment for access copies. Only four (14%) stated that they did not maintain the relationship between access and preservation copies (a fifth respondent did not understand the question). Unfortunately, sites were not asked to explain why this relationship was not maintained. One volunteered that access was handled in a separate system, via export from the repository.

The question of how the relationship between access and preservation copies was maintained in the repository was further explored in the follow-up interviews. Interviewees were asked: *do you record relationships between files, for example, different versions of the same file? How do you record/express these relationships? (For example, through metadata, file names, directory structures, other)?*

Responses from both the interviews and the surveys varied widely. Several indicated the use of an external resource discovery system (e.g., an online catalog) to maintain these relationships. One is using a vendor-provided digital asset management system. Some respondents are still in an investigatory phase. Some noted that where access copies are generated from preservation copies, no relationships are recorded because none are needed. Other respondents indicated the use of metadata, XML schemas, directory structure, file naming conventions, relational tables, identifiers, and manual processes. The following slightly edited responses reveal some of the details, and potential complexity, involved in the various strategies that were reported.

4.2.1 Identifiers

> We use URN:NBN as a naming convention for all files and packages that are stored within the preservation archive. To keep track of copies we use additional numerical codes. Within each package are also stored metadata in a conformant format (XML according to our own XML document format) and in the metadata we can express the relationships between manifestations. It is a practical solution for us at this point. We think it gives us a possibility to migrate later the entire archive to an XML database or some other solution if it will be more useful.
>
> Each storage media object is numbered using a scheme capable of relating redundant versions and making it possible to perform keyword searching in the future in order to easily identify all storage and access media objects for a given collection, for example. Typical values might include record unit, accession number, folder number, and object in series number.
>
> Masters and derivatives are differentiated through the use of role codes in their persistent identifiers. Each item within the repository is assigned a unique persistent identifier (PI), by which it is named and located and which, being hierarchical, can be used to indicate the relationships between objects, their roles (master, view copy) and generation versions. Relationships between objects are also maintained within the management system metadata, but in general may also be deduced from the PI for the object alone.
>
> The relationship between preservation masters and their access instances is maintained by a naming scheme that assigns identifiers to objects that identify their role within the repository thus linking all instances of an object together within the repository, e.g. “12345_pm” for a Preservation Master, “12345_as” for the Access Source copy of that preservation master.

4.2.2 Directory Structure / Filenaming Conventions

> [We] keep related copies/files together by storing each item submitted on its own CD-ROM. If there are multiple copies/files for each sub-item within the item (e.g. from conversion to text and Unicode) then these are placed in separate folders on the CD-ROM. In some instances, the files may include scanned TIFF images of a few pages of the text to help researchers who are studying an uncommon alphabet see the proper formation of the letters. Relationships are recorded/expressed through directory structures. Also, the audit trail is in the metadata record, which is stored on the CD-ROM with the files.

4.2.3 Metadata

Metadata are used to record whether a link could be resolved successfully or not.

We record in relationship metadata structural dependencies (e.g., is part of), derivative relationships, conforms to type (e.g., is DTD for), auxiliary or processing files that are format- or vendor-specific (e.g., is target for, is ICC profile for, is waveform reduction file for). Auxiliary files may be used to relate objects supplied by two different depositors, but these auxiliary files will be generated by someone other than the owner of the archive. We also capture derivative relationships. What we will need to do but don't do right now is "this was derived from that through something else that no longer exists," e.g., a JPEG2000 is derived from a JPEG which was derived from a TIFF, but then the JPEG goes away so what is left is the TIFF and the JPEG2000. These relationships are recorded through metadata.

Metadata associated with the master-derivative relationship includes identification of the master copy, enumeration of what derivations have been produced, and by what methodologies.

4.2.4 XML

Investigating doing this through METS schema.

The repository uses METS documents for the structural information, and a proprietary database for container identification and versioning maintains relationship between access and preservation copies.

4.2.5 Tables

A relationship table records structural and derivative relationships between files. This includes the identifiers of the two related files, the type of relationship, and the identifier of the event creating or associated with the relationship.

The same file may be a component of more than one logical object. For this reason, whole/part relationships are also expressed in tables, to ensure that a file cannot be deleted when it is no longer needed in one context if it is still needed in another.

4.2.6 Mixed strategies

Relationships among files may be recorded through metadata, file naming conventions, and directory structures.

We are not currently committed to only one option here. Relationships may be kept by either the DOM system or external resource discovery system; this has not yet been finalized.

In one case relationship links are made from a catalog to specific electronic files. In another, relational links are made through relational database links.

Database tables, METS files, and occasionally Dublin Core metadata elements are used.

> Relationships are indicated several ways: a) explicitly with an attribute, for example "this is a reference version," b) based on file format and size, c) within the file name. The naming convention includes an institutional identifier, accession number, sequence number, file number, file format and extension.

Some institutions discussed relationships between objects that went well beyond that between access and preservation copies:

> One other relationship recorded has to do with whether the object in the repository is "duplicated" in the library collection: for example, a rare print book in the library collection that has been digitized, with the digital version residing in the repository. Other relationships that are recorded already in the repository include: a) between master images (in the gallery, photo lab, xray), b) between master images and original work of art, c) between components of a complex object (an artist book, sculpture, installation piece, etc.), d) between parts and whole objects. The institution uses a combination of EAD and METS to define all aspects of the collections from the collection level through to the individual components of complex objects.

> The organization currently deals with a complex descriptive system for existing archival objects. They will want to adapt this system to digital records as well, but have not determined how. They anticipate maintaining the kinds of relationships that the current analogue systems take into consideration, such as relationships between records and who created them (individual, agency), and relationships between records (are the records part of the same digital asset management system, or used to provide access to other records). They anticipate recording relationships about multiple copies held on different media (film, microfilm) or on the same media (part of their preservation strategy will include keeping multiple copies at the same level in case of corruption or damage, etc.). The hierarchies of relationships that already exist might be used but in a different format in the preservation repository. Currently, with the physical archives, relationships are expressed in finding aids (external to records). With the preservation repository, there will be more need to put more of the relationships into the metadata and they have not decided on a strategy for this yet.

Interviewees were asked specifically about their treatment of intellectual relationships, as opposed to relationships between files: *Do you record intellectual relationships between objects, for example, objects that are part of the same collection, or objects that are different editions of other objects?*

Responses ran from “no” through “yes,” with a range in between. Some of the intermediate responses indicated that these relationships might be inferred from other metadata, in the form of billing or project codes. Others indicated that original representations of these relationships would be preserved, or would be recorded in descriptive metadata, or would be reflected in the organization of the deposited material.

> Generally no intellectual relationships are recorded. The only way that that might be captured, very indirectly, is through project-related billing codes in some cases.

Metadata for each file includes a link to the metadata for the intellectual entity of which the file is part.

Yes, in many ways. Content files are aggregated into "bundles" of related material that belong to logical “items” and which are further aggregated into "collections" (and items can belong to multiple collections). Versions can also be accommodated via multiple item bundles or via metadata to link separate items depending on the versioning relationship (i.e., is it multiple versions of a document over time, or multiple versions of a work in different formats, or a thumbnail, reference copy, and archival master of the same image).

4.3 How are metadata and materials stored within the preservation repository? For example, one repository might zip metadata together with content files and store the zip file as a single entity. Another repository might store metadata in relational database tables and content files as individual entities in a file system. A third repository might use multiple models for different types of materials. Please describe all of the models that apply.

Seven respondents provided no answer to this question, or indicated that the storage architecture for the repository was not yet determined.

There were four common models:

1. Content data objects are stored in a filesystem; metadata are stored in a relational database management system. (Fourteen responses.) In a variation on this, data are stored offline (e.g. on DVD or CDROM) and the metadata are stored in a database. (Two responses.)
2. Content data objects are stored in a filesystem; metadata are stored in an XML database. (Three responses.)
3. Content data objects are stored in a filesystem; metadata are stored in an RDBMS; full copies of the metadata are also stored with the objects in the filesystem. (Three responses.) In a variation of this, the data and metadata are stored on media (CDROM) and the metadata replicated in a database or spreadsheet. (One response.)
4. Content data objects are stored in a filesystem; some metadata are stored in an RDBMS; other metadata are stored with the objects in the filesystem. (Five responses.)

The largest number of respondents had a simple architecture in which the metadata were stored in a database while the content data objects were stored in a filesystem (models 1 and 2). However, several of these noted they were evaluating or planning to move to a different architecture. Three noted they were considering model 3, in which metadata are also replicated with the content data objects. Two were users of DSpace who noted the application was considering use of stored METS files for some metadata currently stored in the database. One was looking at storing objects in the same database as the metadata.

Respondents implementing model 3, where metadata are stored in two places, may have done so differently. One said the metadata would probably be stored as an XML file treated as a content data file. One was zipping the metadata and content files together. A third respondent noted that both metadata and content files would be stored in the same container. There is no way to tell whether these are three different methods or different ways of describing the same method.

In model 4, some metadata are stored with the object and some separately, depending on the type of metadata. How this was divided varied. One repository planned to store "unchanging" metadata with the object, but did not elaborate on what that was. Another repository stored preservation metadata in a database and other metadata "in XML wrappers around the content," but "preservation" metadata were not further defined. Another repository was storing fixity and reference (descriptive) metadata with the object, and context and provenance metadata in a database. A fourth stored administrative and technical metadata in a database, except some formats, where additional technical metadata were stored with the object. "More detailed technical metadata only of use to, for example, an audio engineer during the preservation process may be packaged [with the object] in a METS file and be opaque to the [repository management system]."

Other architectures described included:

> Metadata are stored with the objects in a filesystem, but selected metadata are replicated in a database for fast access.
>
> Descriptive metadata stored in the file headers of the content data objects; technical metadata in separate specifications documents.
>
> Metadata and content data objects both stored in a database management system.
>
> Metadata and content data objects both stored in a filesystem.

The above summary does not cover structural metadata for compound objects. Most responses did not specifically mention this. When mentioned, structural metadata were most commonly stored as a METS file, either with the object or separately. We cannot tell from the majority of the responses whether structural metadata were considered part of the object or independent metadata.

Some respondents elaborated on the reasons for their storage decisions:

> All metadata except structural metadata are stored in relational tables to facilitate data access and use. Specifically, we need to be able to retrieve and report on classes of objects selected by many different elements of administrative and technical metadata.
>
> We chose to separate the metadata from the content files for reasons of flexibility and maintenance. This way we can add information without having to retrieve all objects from the storage database.

The strategy for metadata storage involves the separation of metadata management from object management with metadata and content being stored separately (usually) in relational databases and until ingested into the [repository]. Once in the [repository], metadata are stored in XML as objects in the [repository], and the content objects are stored in a number of bitstream managed environments currently using EMC's Centera and tape systems for purposes of file redundancy, efficiency and safety in storage management.

Metadata of which the repository is aware are those characteristics which we feel we need to manage the objects in the aggregate, or to manage individual objects, or to characterize materials for subsequent analysis for preservation action, and also for reporting. So not every piece of metadata we can think of is recorded. More metadata can go into files, but the repository won't manage it. We do it in relational tables, because of scale, control over possible values ranges, and data integrity.

For us, data and metadata are both digital assets, so it is important to preserve both data and all pertinent metadata. ... The canonical forms of our metadata are kept in XML. So far a copy has been kept on disk together with the master files. However, we plan to move the metadata into an XML database, leaving the data files on disk, with the metadata pointing back out to the data files, both master files and service copies. We will not move the data files into the database, because of storage and retrieval concerns.

4.4 What preservation strategies are your preservation repository implementing now? (Check all that apply.) Why did you chose these particular strategies?

Strategies	Count (survey)	Percentage
Restrictions on submissions	20	42%
Making print or microform copies	5	10%
Bit-level preservation	41	85%
Normalization	19	40%
Migration	16	33%
Migration on demand	10	21%
Emulation	5	10%
Other	5	10%

The following question (4.5) asks which strategies are planned for the future. The subgroup suspected that many repositories were not at the point of implementing planned strategies, so these questions were designed to get a picture of present state and future trends. However, answers to 4.4 were unreliable. First, it is unlikely that respondents who indicated they were entirely in planning or development stages could have actually implemented any of these strategies, so numbers were corrected to include only repositories in some stage of production. Second, some respondents appeared to misunderstand the capabilities of their own systems. A respondent using

DSpace, for example, checked “migration” as a currently implemented strategy, while DSpace in the winter of 2003/2004 did not contain any migration capability.

The counts adjusted for identifiable errors or improbabilities, appear in the table below.

Strategies	Count (adjusted)	Percentage
Restrictions on submissions	13	27%
Making print or microform copies	1	2%
Bit-level preservation	21	42%
Normalization	9	19%
Migration	10	21%
Migration on demand	9	19%
Emulation	5	10%
Other	4	8%

Five institutions did not check any strategy at all, indicating they were in planning stages or had not yet determined a strategy. Every repository in production checked bit-level preservation. However, comments indicate that “bit-level preservation” has two meanings. It could mean that objects stored within the repository are protected from unauthorized change, or it could mean that objects as originally submitted to the repository are carried unaltered indefinitely, regardless of other preservation actions taken. We do not know the dominant meaning among respondents.

Restriction on submissions and normalization are both ways to control the variety of formats a repository manages and to simultaneously control costs. Migration and migration-on-demand strategies, while described as “practical” or “cost effective,” are less often used. The fewest institutions make analog copies or use emulation.

Respondents checking “Other” listed technology preservation, the UVC (Universal Virtual Computer) developed by Raymond Lorie (Lorie 2002), and a digital ontology, defined as a description of the structure of the digital entity:

> The document is retained in its original encoding format. The structures in the digital entity, the semantic labels applied to the structures, and the operations that are performed upon the structures are characterized as a set of relationships imposed on the digital entity. The relationships are organized into an ontology to specify the order in which they must be applied, and the mappings to the presentation structures that are desired. The ontology is migrated to new relationship encoding formats over time as the standards evolve.

Only one respondent in production checked a single strategy, and that institution planned additional strategies in the future. As a national archives wrote, “We are taking a practical stance – we are not advocates for a ‘one true way’ of archiving.” Clearly, the vast majority of institutions feel there are good reasons to have more than one approach, if only to hedge their bets in a developing field.

In answer to the question why these particular strategies were chosen, respondents’ explanations show that institutions are making the best decisions they can now with an

eye toward the future. For some, the current strategies are seen as interim or "medium-term" because the preservation path of the file format is unclear, or because there is an assumption that there will be additional technical options in the future. Sometimes, strategies are selected even though the approach raises questions. One national archive noted that migration raises "questions of authenticity and in what dimension…" It is possible that more than one migration would be performed at the same time, to preserve different characteristics of a record."

Previous experience with a strategy (such as bit preservation or migration) was seldom stated as a reason for selecting a strategy. Cost was mentioned, but does not seem to be an overriding reason. Strategies appear to be most often selected to meet access goals or requirements, because of file format requirements, because they allow flexibility in the future, or because they are available.

One national library suggested some reasons why particular strategies were NOT selected, rather than why they were. The library suggested that normalization may work for text, but may be too lossy for audio and video, while emulation lacks tools and is not developed enough to be a practical choice. They concluded:

> Any given strategy depends upon reliable tools with which to implement that strategy. Without tools the minimum preservation strategy for an object is firstly to hold some form of digital "original" in anticipation of the future availability of tools and secondly have an "original" iteration of a digitally born object to act as a "reference copy" for use with those tools.

4.4.1 General (comment could not be associated with a particular strategy)

- The in-house tools are available.
- File format dictates strategy. The file format is pervasive in the marketplace, well supported, or stable.
- Strategies follow current procedures and workflows.
- Access issues: guarantees of future access, ease of delivery, ability to make high-quality duplicates.
- Constraints of associated system: e.g. an access system; a publishing system; future applications that will provide additional functionality for the content, or parse and manipulate the content.
- Quality (undefined) to ensure best quality for long-term preservation of accessioned material.
- Meets business needs of organization.
- Protects the "integrity of original".
- Research project.
- Acceptable standard; a best practice.
- Cost effective. Possible with available resources.
- Simplest way.
- Experience with the process.

4.4.2 Restrictions on submissions

- Based on control at creation, for example for digital surrogates created as part of an in-house digitizing program. Allows control format and quality.
- File format well supported, "standard" file formats (e.g. Tiff, bwf).

- Established relationship with subscribers who conform to our standards.

4.4.3 Making print or microform copies

- Selection policy drives strategy. Example: as some serial paper publications have moved to Web-only publication, we made print copies on archival quality paper for the library's collection.
- Only documentation is printed on paper.
- Existing process for existing reasons. Examples: supply copies to the Library of Congress; film is the juridical (legal) copy of the item.

4.4.4 Bit-level preservation

- Cost, monetary and other resources.
- Flexibility: allows for future strategies, including digital archeology.
- Homogeneous strategy (applies to all objects in archive).
- Have to do this at a minimum.
- Doesn't restrict types of objects to be ingested.
- Practical. Available now.
- Object remains in original form; "it's the most authentic record we have."
- Less intervention by archive on objects ingested.
- Minimizes the risk of data loss or corruption.
- Robust (disaster recovery processes, duplicates, etc.).

4.4.5 Normalization

- Don't need to preserve look and feel.
- Reduce the chances of losing access capabilities through hardware and software obsolescence.
- In the case of databases, experience has taught us that data separated from software and hardware constraints allow researchers to use more modern data analysis software.
- Reduce future migrations because of choosing stable normalized formats and having fewer formats to manage.
- Practical.
- Cost effective.
- Minimizes the risk of data loss or corruption.
- Ease of data management for example extraction of metadata, fewer number of different file formats to identify and manage.

4.4.6 Migration

- Preserves access.
- Have to do this at a minimum.
- Experience with the process.

4.4.7 Migration-on-demand

- Practical.
- Reduces cost.
- No need to restrict submissions.

4.4.8 Emulation

- Extend access over time.

4.5 What preservation strategies are your preservation repository planning to implement in the future? (Check all that apply.) Why did you chose these particular strategies?

Strategies	Count	Percentage
Restrictions on submission	18	38%
Normalization	23	48%
Migration	32	67%
Migration on demand	14	29%
Emulation	10	21%
Other	7	15%

Nine institutions did not check any strategy, some commenting that they are in planning stages, or had not yet determined a strategy. One institution explained that "none of the formats we take are approaching obsolescence, so it's not urgent." Only two institutions checked a single strategy.

The percentage of institutions selecting migration triples when institutions think about future actions. Correspondingly, there are more and different explanations given for making the future choice than for the present (see below). As in question 4.4, restrictions on submissions and normalization remain more popular than migration-on-demand, emulation, or "Other."

Respondents checking "Other" described new approaches, in addition to the Universal Virtual Computer and digital ontology mentioned in response to 4.4. One institution wrote, "We are beginning to archive Web video, a wholly new medium for us. We will probably do some kind of normalization-tape combination." A museum wrote that they are creating "'scores for born-digital art works; a set of instructions that allow the work to be re-created using contemporary technologies if applicable and if allowed". Another institution is investigating the preservation layer method described during the NEDLIB project.

A few institutions indicated that their reasons for selecting certain strategies in the future are identical to their reasons in the present. However, there is little overlap between the reasons specifically associated with a strategy in answers to 4.5 and 4.4. The explanations in 4.5 emphasize the impact of object types, delivery options, the perceived value of the object, and future preservation actions. Cost remains a minimal consideration, possibly because future costs are unknown. Two institutions estimated the cost of emulation at opposite ends of the spectrum, one stating that is may be prohibitively expensive, and the other stating that it may be cheaper than migration.

A few respondents voiced concerns about future preservation actions. One issue is the possible cost of taking action and whether the action is feasible without automated processes on large number of objects. When considering migration-on-demand, there are concerns about scale and the number of transformation programs that would have to

be maintained over time. Some of the strategies chosen raise issues or questions when compared with the institution's mission, mandate, user needs, or definition of authenticity. For example, restricting the types of formats is not possible in institutions with a national deposit mandate. Some institutions will not be able to restrict the types of content they receive. It was pointed out "normalizing and migration have their limitations in a preservation sense in that an object may ultimately become so removed from its original self as to loose all relevant, or desirable 'look and feel.'"

Clearly, the future is unknown: "We are keeping an open view on all approaches and, in light of outcomes of research, will implement whatever pathways prove to be appropriate for the range of materials in the repository (including "Other" pathways which may evolve in the future)." Another institution echoed that opinion: "Our operative principle is that we know so little about long-term preservation, the more versions we keep around the more likely we'll be able to continue transforming the content into something usable. At some point it will become too expensive to maintain so many versions, but hopefully by then we'll know more and be able to make more informed decisions about what to keep." A third institution indicated that the only way to move forward is by trying different approaches: "Only by engaging with material and by learning from experience can we develop more comprehensive strategies and policies." A suitable closing remark is, "In order to respond to a dynamic and changing external environment any chosen preservation strategy must itself be flexible and dynamic if it is to be an adequate and enduring response."

4.5.1 General (comment could not be associated with a particular strategy)

- The in-house tools are available.
- File formats dictate strategy.
- We expect to deal with projects in the future that insist on a particular encoding format.
- Access issues: guarantees of future access.
- Flexibility for future choices and multiple file formats and publications.
- Protects the integrity of original.
- Best practice.
- Cost effective. Possible with available resources.
- Practical; seems feasible given the current state of technology.
- Experience with the process.
- Influences of version and designated community.
- To better manage content.

4.5.2 Restrictions

- Not so many file formats that need to be handled.
- Can select "archivally appropriate" formats.
- Standardization of formats and quality.
- Based on whether preservation is guaranteed for an object.
- Cost-effective.
- Control over metadata.

4.5.3 Normalization

- Optimise for delivery.
- To simplify and reduce future migrations or other preservation actions.

- Determined by object type, e.g. Complex objects.
- Determined by format, e.g. Sgml or xml with dtds/schemas.

4.5.4 Migration

- Easy.
- Based on value of object, e.g. “for critical objects”.
- Based on object type, e.g. Documents.
- Optimize for future preservation actions.
- Optimize for enhanced delivery.
- Possible to do batches of objects.

4.5.5 Migration-on-demand

- Based on users’ needs.
- Based on value of object, e.g. “enduring value”.
- It is only as an alternative strategy that migration will be considered.

4.5.6 Emulation

- Based on object type, e.g. “certain born -digital art works”.
- Based on cost “Emulation – might be not as expensive as migration, but difficult”.
- Authenticity, e.g. to maintain original format.

4.6 What type of applications software is in use in your preservation repository? (Check all that apply.)

Applications	Count	Percentage
Commercially available software	25	52%
OpenSource software	26	54%
Locally developed software	33	69%

Eleven institutions selected only one category (six local, three commercial, two open source) but most selected multiple categories. It was clear that, to some extent, the answers depended on the respondent’s interpretation. For example, one checked “locally developed” only, but explained in a comment that the local application used a commercial database. Another with the same situation checked both “commercially available” and “locally developed.” Overall, it is clear that most institutions are building digital preservation systems from a variety of components in all categories.

Thirty-three respondents indicated use of some locally developed software. Three institutions described local development of complete digital preservation systems. In most cases, local components were developed as pieces of a larger commercial or opensource system. Examples of these include:

- Enhancement of the TKL software with an ingest client.
- Preservation metadata extract tool. (Two responses.)
- Various interfaces for clients. (Two responses.)

By far, the majority of respondents used a combination of types of software applications. Seventy different commercial and open source software products were specifically named. Of those, only eleven were named by more than one respondent, and most of these were system software or database tools rather than repository applications. The products named by more than one respondent are listed below with the number of respondents in parenthesis.

DSpace	9
Oracle [various pieces]	8
Apache	4
Fedora	4
OCLC Digital Archive	3
Unix	3
DB2	2
LINUX	2
MySQL	2
PostgreSQL	2
Tivoli storage manager	2

The overlap in system software is probably underreported; most respondents did not list specific webservers, operating systems and databases although they are likely to be using these products. There is little standardization on digital repository systems. The software product mentioned most often, DSpace, seems to be most popular among North American research libraries. Only two of the nine institutions using or evaluating DSpace were not North American research libraries; one was an American archive and the other a national library that was also evaluating Fedora.

4.6.1 Databases

Commercial	**Local or Open source**
DB2	MySQL
Filemaker Pro	PostgreSQL
Informix	
MS Access	
MS SQL Server	
Oracle 9i	
Sybase	
Tamino (Software AG)	

4.6.2 Digital Repositories / Preservation Systems

Commercial	**Local or Open Source**
DIAS (IBM)	DSpace
OCLC Digital Archive	DAITSS
	GNU eprints

4.6.3 Digital Library/Digital Content Management Systems

Commercial	Local or Open Source
CONTENTdm	CourseWork
DigiTool (ExLibris)	Sakai tools
ENCompass for Digital Collections (Endeavor)	TKL (The Keystone Library)
Insight (Luna)	DLXS (University of Michigan)
TEAMS (Artesia)	Fedora
SIRSI Unicorn	Greenstone
Virage	

4.6.4 Harvesters

Commercial	Local or Open Source
	HTTrack
	NEDLIB
	wget

4.6.5 Storage systems and file managers

Commercial	Local or Open Source
Atlas data store	LUSTRE
ASM from StorageTeK	Opus (GPSoftware)
Castor	Resource Storage Broker
Centera (EMC) for storage management	
DCache	
DMF	
HPSS	
Tivoli Storage Manager / ADSM	
Unitree	
Veritas (Sun)	

4.6.6 Miscellaneous Tools

Commercial	Local or Open Source
Autonomy	Checksum checkers (misc.)
Bscan	HTML-Kit
Legato	JHOVE
Semantic Information Router (Profium)	JPedal PDF parser
Virus checkers (misc.)	Tiday.exe

5 Metadata

5.1 What categories of metadata are (or will be) stored by and used by your preservation repository? (Check all that apply.)

Categories	Count	Percentage
Rights and permissions	37	77%
Provenance (document history)	40	83%
Technical metadata	41	85%
Administrative and management information	41	85%
Bibliographic/descriptive	38	79%
Structural metadata	36	75%
Other	9	19%

Most respondents are recording a wide range of metadata. Eight institutions checked all categories; twenty-five institutions checked all categories except "Other." Nine institutions checked "Other," with five giving explanations. Three described these metadata, at least in part, as an audit log of preservation actions applied to data objects which would probably be considered digital provenance. Others mentioned behaviors, annotations, semantic mark-up of the content, versions, containers, and checksums. The last three could be considered provenance, structural, and technical metadata respectively.

Some respondents editorialized about preservation metadata. For example, some institutions commented that rights and permissions and bibliographic/descriptive metadata were not preservation metadata. Others asserted that audit trails were preservation metadata. One national library respondent, while asserting the need for "a clear definition" and "demarcation lines" around preservation metadata, concluded that "preservation as an activity will rely on all metadata and not just that deemed to be unique to preservation metadata and that all metadata will be available in a unified and meaningful way for all processes, e.g. resource discovery as well as preservation."

Some institutions have guidelines as to when metadata are applied. Examples: "The only document history information we record begins with the ingestion of the object; we do not record prior history." Or, metadata are limited: "The archive stores only minimal bibliographic information (identifier, title, and for serials volume and number)." Other metadata are limited to a specific format, for example, one institution retains provenance only for photographs. Some metadata may not be needed because of the institution's relationship with the depositors. "We do not record rights metadata because that is the responsibility of the submitting library; their Agreement with the archive states [the archive has] the right to view, copy and make derivative versions of any material they submit, and that they grant these rights to the [archive.]"

5.2 If you are using or planning to use metadata elements from one or more published scheme, which schemes are you using? (Check all that apply.)

Schema	Count	Percentage
AUDIOMD: Audio Technical Metadata Extension Schema	3	6%
CEDARS	7	15%
Creative Commons Metadata	5	10%
METS	26	54%
MIX or Z39.87	12	25%
MPEG7	1	2%
MPEG21	4	8%
NEDLIB	7	15%
National Library of Australia	6	13%
National Library of New Zealand	7	15%
OCLC Digital Archive Metadata	11	23%
TEXTMD: Schema for technical metadata for text	7	15%
Schema for rights declaration (METSRights.xsd)	5	10%
VERS	2	4%
VIDEOMD: Video Technical Metadata Extension Schema	3	6%
Other	22	46%

Eleven respondents did not check any of these categories. Four of these indicated they had not yet determined what they would use. One explained that they had developed their own set of elements after reviewing existing metadata schemas, but they expected to use existing schemas for technical metadata. Others listed METS, MPEG 7, and the schema for rights declaration as being under consideration.

METS was by far the most commonly used scheme. Survey results indicate adoption in all three types of institution, although to varying degrees: 64% of libraries, 42% of archives, and 35% of other institutions used or planned to use METS. Z39.87, or its XML representation MIX, was the most widely used technical metadata scheme, but its use was almost entirely within the library community (ten libraries and one "Other"). This could be because libraries are most likely to archive still images, or because as a NISO standard Z39.87 has not been promoted in other communities.

A local metadata scheme based on other existing schemes is common, as is the use of more than one scheme in the repository. Some respondents mentioned mapping from a local metadata scheme to a standard scheme, although only one explained why: "We're faced with mapping more than two dozen nonstandard metadata sets from different databases … and are planning to organize them through a DCMES [Dublin Core] crosswalk." For others, the purpose may be to discover the completeness of the local scheme through comparison with something else. The following comment from a national library is typical of the blending underway in some repositories:

> Existing metadata in the Digital Collections Manager handling digital surrogate images draws on elements from Z39.87 (NISO Technical Metadata for Digital

Still Images), and is capable of being mapped to METS for metadata transmission. We are in the process of further determining our specific preservation metadata requirements, developing a data model and seeking ways to implement collection of the required preservation metadata within the repository management systems. We are discussing and seeking to implement the RLG/OCLC recommendations for preservation metadata, which draw on the CEDARS, NEDLIB and (draft) NLA schemes. We do not expect to exclusively or explicitly use any one of the above schemes, but to adapt recommendations and elements to fit our determined requirements for particular materials, actions to be managed, business processes and system capabilities. We would expect to use an output model to populate elements for each entity on demand, mapping from existing (or extractable) metadata from anywhere in the system.

Fifteen respondents marked "Other" along with one or more of the given schemes, and seven marked only "Other." Many of the other schemes mentioned were descriptive metadata, because the survey instrument did not include primarily descriptive schemes in the list provided. Specified schemes included:

- Anglo American Cataloguing Rules (AACRII)
- Audio Engineering Society's drafts for Core Audio technical metadata
- CERA The Climate and Environmental data Retrieval and Archiving (http://www.mad.zmaw.de/Services/Lectures/CERA@Kiel.pdf)
- CF conventions for climate and forecast metadata (http://www.cgd.ucar.edu/cms/eaton/cf-metadata/)
- Data Documentation Initiative (DDI) (http://www.icpsr.umich.edu/DDI/)
- Dublin Core, or Dublin core based (7 institutions)
- EAD (2 institutions)
- MAB2 scheme, Maschinelles Austauschformat für Bibliotheken (Automated Library Exchange Format)
- MARC and MARCXML
- NewsML (an XML-based metadata schema promulgated by the International Press Telecommunications Committee and the Newspaper Association of America (NAA).
- ONIX
- Record of Archival Description (RAD)
- SMIL (Synchronized Multimedia Integration Language)
- Spectrum
- TEI Header
- UNIMARC
- VRA Core
- WAGILS modified to include administrative and preservation elements

5.3 Does your repository record information about these types of entities? (Check all that apply.)

Types of Entities	Count	Percentage
Collection	33	69%
Logical object such as a book or photograph	41	85%
Non-digital source object	23	48%
File	41	85%
Bitstream	24	50%
Metadata	26	54%
Other	10	21%

The use of metadata at various levels can be summed up in the words of one respondent: "We do create (or expect to create) metadata at all these levels[;]…not all levels will be relevant for all materials, and the required details at each level are not yet fully determined. Wherever we have an entity to manage we would create metadata." Most respondents checked multiple categories, and most institutions record metadata about both logical entities (collections and/or logical objects) and about digital entities (files and/or bitstreams). Although three respondents claimed to record only file or bitstream metadata in the preservation repository, one explained they keep collection and logical object file metadata in external databases.

Because of the high number of respondents in each category, it is interesting to look at which institutions did NOT select certain categories. Three respondents did not answer this question at all, presumably because their metadata structures were still under development. Eleven institutions do not record information at the collection level, including six national or state libraries. The proportion of national and state libraries that did not record collection level information was more than twice that for other types of libraries, although the size of the sample is too small for this to be significant. A comment by one respondent hinted that these libraries might consider collection level information the purview of the library catalog or other databases external to the repository.

Four institutions do not record information about the logical object. All of these institutions record metadata at the file level; one of them also records information at the collection level. Description of non-digital source materials was considered by respondents to be either links to descriptive metadata, or the descriptive metadata itself. Examples of the former included a link to a bibliographic record describing the original, and a catalog key. Examples of the latter were: standard numbers, creator, title, publication information, subjects, and source type.

All but four respondents checked that they recorded file level metadata. However, one of them uses the OCLC Digital Archive, which creates metadata for the institution at the file level once the object is ingested. Another uses DSpace, which captures some file level information even if not provided by the depositor. Correcting for this, only two institutions do not record file level metadata.

The survey question defined the bitstream by stating it "may be equivalent to a file, a subset of a file such as a binary object embedded in a PDF, or greater than a file, such

as a digital video stored in three parts." Half of the responding institutions record metadata at the bitstream level. Only one institution did not record metadata at either the file or the bitstream level; this was an "Other" (for-profit) organization that recorded only collection and logical object level description. This repository appears to rely upon its own knowledge of various file types to do data migrations. In general, the preservation community considers technical metadata pertaining to files and/or bitstreams essential for preservation activities such as migration and emulation.

Respondents were asked to provide some examples of metadata they supplied at each level. Replies included (this list is not inclusive):

Collection

- Free-text description
- Link to a bibliographic record describing the collection
- URL
- Related objects, e.g. re-harvests of the same URL
- Collection name
- Resource type
- Series information
- Selector, project name, project coordinator, digitization and work plan documents

Logical object

- Title
- Contributor, contributor email
- Relationship element to tie all the files to the logical object
- When it was harvested
- Related logical object level metadata, such as a bib record number
- One pre-existing identifier known to the outside world, and minimal bibliographic information (title, serial volume, serial issue)
- Copyright authorization
- Expiry date of content
- Related documents e.g. earlier or later versions

File

- Purpose of file (access, preservation etc.)
- Create date, date/time of last copy, date of last file modification
- A code indicating if the file is a submitted original or a normalized or migrated version
- Whether the file contains embedded links to other files
- Access inhibitors such as encryption
- Any problems or abnormalities noted about the file
- Relationships for the file
- Layer level
- Number of bytes downloaded and byte order
- Evidence of any redirections in the process of resolving the URL
- Character set encoding used in the file

Metadata

- Creation and modification dates
- Creator of metadata
- Schemas used, schema versions

5.4 How is metadata obtained (or expected to be obtained) by the preservation repository? For example, is it submitted by depositors, extracted automatically by the repository's computer programs, other? If different methods are used for different sets of metadata, please note all of them.

Method of Obtaining	Libraries	Archives	Other	Total	Percentage
Supplied by depositor	22	6	8	36	75%
Extracted by program	20	6	10	36	75%
Supplied by repository staff	16	6	8	30	63%
Other				34	71%

These results indicate that institutions are using a combination of techniques, dictated by the type of metadata. Trends include:

> When individuals were supplying the metadata, the process was by manual input. When publishers did the supplying, it was by batch load.
>
> Automatic extraction by the repository software was often limited to technical metadata. In at least two cases, it was used to build descriptive metadata from full text content, although it was noted in one of these that accuracy was suspect and needed to be checked by institution staff.
>
> When the institution supplied metadata, it was usually entered manually by staff (57%), and to a lesser extent extracted or derived from other bibliographic databases (23%).

Many survey respondents expressed a desire and expectation that metadata creation, especially administrative and technical metadata, would be automated as fully as possible in the future. In follow-up interviews, institutions that indicated their repositories did extract metadata automatically were asked for more detail about the automatically supplied metadata elements. In all cases, technical metadata were extracted from file headers. One interviewee obtained structural metadata, and one obtained descriptive. The second respondent mentioned that descriptive metadata are often suspect, and that they are not comfortable with reliance on tools that automatically extract descriptive metadata.

Interviewees were asked, if they obtained metadata both from the submitter and by automatic extraction, could this result in duplication of information and if so, how they reconciled differences. Four of thirteen respondents said this was possible. Two reviewed such conflicts on a case-by-base basis. Of the other two, one always favored their own extracted values, and the other one always favored the submitted values.

5.5 Do you require contributors to your preservation repository to provide metadata with their contributions? If so, what are your requirements? Please provide sample documentation if possible.

There was some ambiguity in the answers to this question. While 75% said they expected certain metadata to be supplied from submitters, fewer (only 65%) went so far as to say they *required* it. The proportions of institutions requiring metadata were roughly the same among libraries, archives, and other. This might be explained by the nascent state of these preservation repositories and the fact that their policies are still being developed and refined. Also, repositories may be trying to avoid any perception of a hard and fast barrier to deposit, which might be how some submitters would view such a requirement.

The type of metadata being supplied by submitters was predominantly descriptive (64%), with a much smaller portion administrative or technical (22%). Based on the responses, there is a wide variation in the expectations of the repositories, both in terms of quantity and complexity. The range includes "some basic fields" of descriptive metadata to relatively complex technical descriptors and transfer forms.

Fourteen respondents were interviewed about their metadata practices. Those that accept metadata from submitters were asked to describe this in more detail. Answers included a wide range of metadata types: descriptive, technical, structural, and administrative, but the most common were descriptive and technical (in that order). The number of metadata elements accepted also varied, from minimal (only a few fields) to extensive, which, in some cases, was fifteen or more fields. One respondent described its own range as from "the barest details to a solid cataloging record." Several repositories tailored the elements to the specific access needs of the submitter. Vendor-supplied metadata were usually batch-loaded, the result of careful planning with the repository staff. Government repositories generally require more extensive information, similar to their requirements for non-digital deposits. But in most cases, the type and quantity of submitted metadata are the result of a consultation process between repository staff and submitters.

Interviewees were asked if they had, or planned to have, any process in place for assuring or checking the accuracy of submitter-supplied metadata. Two-thirds replied affirmatively to this, although there was a wide range of techniques within that group. Nearly all of the affirmatives included one or more passive feature such as providing clearly articulated standards for submitters, offering assistance when requested by submitters, and including features such as drop-down menus for metadata capture. Far fewer said they would assign staff to perform basic quality control and make corrections, and only one said they would go so far as to have their staff perform in-depth quality control and upgrade of the submitted metadata. In that single case the respondent said that in-depth quality control "may be done for descriptive metadata in certain categories," such as when MARC is available, "but this may not be scaleable in the future."

In sum, metadata requirements are driven by a number of factors influencing what metadata to record and how to capture or create them. Respondents indicated that they considered several issues in determining which metadata to collect. They considered which metadata end users from different constituencies needed to access content that may vary across subject disciplines; which metadata the repository needed to manage

the preservation of content; which metadata would be available from the content creators; and which metadata are suggested by the evolving standards work in the international community. Respondents also considered how the metadata would be created or collected. Cost is certainly a factor; some respondents said that they wanted to reuse existing metadata if possible, or create them automatically.

5.6 How is metadata stored and updated in your preservation repository? If multiple methods are used, please explain.

Storing Method	Count	Percentage
In a relational database	33	69%
Bundled with related content files	22	46%
In an XML database	13	27%
In a proprietary database or format	11	23%
In flat files	9	19%
In an object-oriented database	2	4%

Libraries and archives may already have a significant investment in relational database technology that is being deployed for a variety of tasks including the preservation repository. The significant number of responses for "Bundled with related content files" points to a significant level of acceptance of the OAIS concept of the "information package" among organizations in the sample. Many respondents use some combination of storage techniques, depending on the nature of the material being handled and prevailing practices and environment in the institution. When two or more methods were recorded, it was most common for the data to be stored in both relational tables and bundled with content. This appears to be an emerging best practice for metadata storage.

Number of Methods	Count	Percentage
1 method	9	19%
2 methods	21	44%
3 methods	5	10%
4 or more methods	3	6%
No response	10	21%

5.7 Some preservation repositories create or plan to create normalized or migrated versions of digital materials. If this applies to your repository, how will metadata for the original and new versions be handled? What information (if any) will be recorded only for the version created by the repository?

A majority (twenty-eight) gave no answer or indicated that no approach has been adopted. Of the twenty substantive responses, all but one indicated that multiple versions (originals and subsequent manifestations) would be maintained in the

repository, and that metadata would be created and/or maintained for all versions. Only one specifically indicated that the original versions would not be retained. This strong preference toward retaining the original object points, perhaps, to a degree of uncertainty in this early period of development of preservation systems, and certainly indicates a best practice.

Approaches to metadata were divided between distinct records of various types (descriptive, technical, administrative) for each manifestation or version, and unified metadata records created or adopted on ingest and updated as subsequent manifestations appear or actions are taken. Several respondents indicated that descriptive or bibliographic metadata would remain static and be copied, while technical and administrative metadata would be dynamic or automatically generated. Several respondents indicated the importance of recording relationship/links between versions and their metadata.

To the question of what information would be recorded for migrated versions, thirty-six gave no response, or indicated that no decisions have been made. Among the eleven responding, the following information elements were mentioned twice or more:

- File format/media
- Relationship (between objects and metadata)
- Audit trail/history of conversion/migration/normalization processes

Other elements mentioned were:

- Date
- Purpose of migration/reformatting
- Administration
- Fixity
- Technical consequences of migration

5.8 What metadata do you feel is most important to record for preservation purposes?

Ten respondents gave no answer to this question. Among the thirty-eight that responded, the following elements or concepts were mentioned twice or more:

- File format & version
- Relationship between files/related content/objects
- Descriptive metadata
- Copyright/rights information
- Technical metadata
- Creator
- Administrator of archive/repository/preservation processes
- Date
- Administrative metadata
- Migration/normalization history/processes
- Version history/change history

- Provenance (generally and specifically digital)
- New Zealand metadata scheme
- Audit trails
- Context
- Contributor/submitter/donor

Other elements or concepts mentioned include:

- Provide reference to a format registry
- Identify and characterize
- Document relationships with other objects
- Provenance and change history
- MARC bibliographic information
- Significant properties
- Authenticity
- Access controls
- Relationship between metadata and location of materials
- Preserve context of content
- Attributes of content
- Ingest actions
- Fixity (storage and migration processes)
- Hardware and software requirements

The responses indicate that priority, or emphasis for individual implementations, may grow largely from local context and needs, as well as specific interpretation and understanding of various aspects of preservation and metadata. They also reflect the absence of a unified, commonly agreed-upon lexicon in the field.

6 DISCUSSION

Any conclusions from this survey must be tentative at best. There were only forty-eight completed surveys included in the results. Although this may represent a reasonable proportion of the cultural heritage institutions developing preservation repositories, this is still a very small number. There are other caveats as well. We know of several important repository initiatives that did not respond. Many respondents did not distinguish between what they planned to do, and what they actually are doing. Less than a quarter of respondents had actual production experience in implementing active preservation strategies.

Those institutions engaged in digital preservation activities still lack a common vocabulary and, to a large extent, a common conceptual framework. For example, results for many survey questions were muddied by differing interpretations of terms and phrases, from “dark archive” to “access.” And, although most respondents claimed to have been informed by the OAIS reference model and to be at least partly compliant with it, there was substantial difference of opinion as to the meaning of OAIS compliance. Although OAIS has been praised for providing a standard vocabulary in its glossary of terms, it is clear that most of these terms have not been widely adopted in the community, at least not in informal communications such as survey responses.

In relation to metadata, most respondents were recording several different types of metadata and more than half were recording metadata in all of these categories: rights, provenance, technical, administrative, descriptive, and structural metadata. Repositories appear to draw metadata elements from various schemes to suit their purposes. METS, NISO Z39.87, and the OCLC Digital Archive metadata set were the only named schemes used by more than 20% of respondents. Overall, thirty-three different metadata element sets, or rulesets, were mentioned by at least one repository. In general, the survey shows a picture of a community trying to take advantage of prior work but not at the point of developing or settling on dominant standards.

Nonetheless, there does seem to be some discernable agreement in some areas. As one PREMIS member put it: "It may be premature to identify emerging Best Practices, but there do seem to be some emerging Practices."

- Metadata stored redundantly in an XML or relational database and with the content data objects. Metadata stored in a database allows fast access for use and flexible reporting, while storing it with the object makes the object self-defining outside the context of the preservation repository.

- Use of the METS format for structural metadata and as a container for descriptive and administrative metadata; use of Z39.87/MIX for technical metadata for still images.

- Use of the OAIS model as a framework and starting point for designing the preservation repository, but retention of the flexibility to add functions and services that go beyond the model.

- Maintenance of multiple versions (originals and at least some normalized or migrated versions) in the repository, and storage of complete metadata for all versions. Retention of the original reduces risk in case better preservation treatments become available in the future.

- Choice of multiple strategies for digital preservation. There are good reasons to have more than one approach in a developing field.

- Distinction between types of objects. Preservation repositories record metadata pertaining to many different types of things: collections, logical objects, files, and bitstreams. In normal conversation, many of these entities are simply called "objects." Repository systems might make more granular distinctions and explicitly relate the different metadata elements they record to the appropriate types of entities.

It is interesting to compare conclusions of the PREMIS survey and the CENDI report (CENDI 2004). CENDI focuses on preservation issues and access to scientific information, while PREMIS focuses on repository implementation and metadata issues. Nonetheless, PREMIS and CENDI clearly point at similar emerging practices. Some practices that the CENDI report describe as emerging are clearly visible from the PREMIS results, including the widespread use of METS, the practice of storing metadata

redundantly in a separate database and with the stored objects, and the idea of keeping options open in the choice of preservation strategies.

7 REFERENCES

CENDI 2004. Gail Hodge and Evelyn Frangakis, “Digital Preservation and Permanent Access to Scientific Information: The State of the Practice,” Sponsored by the International Council for Scientific and Technical Information (ICSTI) and CENDI, 2004. Available at http://www.dtic.mil/cendi/publications/04-3dig_preserv.html.

Cervone 2004. H. Frank Cervone, “The Repository Adventure,” Library Journal, June 1, 2004. Available at http://www.libraryjournal.com/article/CA421033.

Flecker 2002. Dale Flecker, “Council on Library and Information Resources Survey on Digital Archiving,” in *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*, 2002. Available at http://www.digitalpreservation.gov/repor/ndiipp_appendix.pdf.

Lorie 2002. Raymond Lorie, “The UVC: A Method for Preserving Digital Documents – Proof of Concept,” 2002. Available at http://www.kb.nl/kb/hrd/dd/dd_onderzoek/reports/4-uvc.pdf.

NDIIPP 2002. *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*, Library of Congress, 2002. Available at http://www.digitalpreservation.gov/repor/ndiipp_plan.pdf.

NSF 2002. “It’s About Time: Research Challenges in Digital Archiving and Long-term Preservation,” Final report of the Workshop on Research Challenges in Digital Archiving and Long-term Preservation, April 12-13, 2002, sponsored by the National Science Foundation and the Library of Congress. Available at http://www.digitalpreservation.gov/repor/NSF_LC_Final_Report.pdf.

NSF 2003. “Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation,” 2003. Available at http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf

APPENDIX A: Alphabetical List of Survey Respondents

Archives New Zealand
Ars Electronica Center
Arts and Humanities Data Service
Austrian National Library
Berkeley Art Museum
Bible for the Future Archive
Bibliotheque Nationale de France
Brigham Young University
British Atmospheric Data Center
British Library
Council for the Central Laboratory of the Research Councils
Courtauld Institute of Art
Die Deutsche Bibliothek
Duke University
Florida Center for Library Automation
Government Printing Office (US GPO)
Harvard University
Hochschulbibliothekszentrum (HBZ)
Helsinki University Library / National Library of Finland
Illinois State Library
Indiana University Library
JSTOR
Koninklijke Bibliotheek
Library and Archives Canada
Los Angeles Times
Massachusetts Institute of Technology
National Archives and Records Administration (NARA)
National Archives of Australia
National Library of Australia
National Library of New Zealand
National Library of Portugal
New Mexico State Library
OCLC
Ohio State University
San Diego Supercomputer Center
Smithsonian Institution Archives
Stanford University
State Library of Ohio
State Library of Tasmania
Swiss National Library
Tolkien Society
UK Data Archive (ESDS)
UK National Archives
University of Calgary
University of Chicago
University of Michigan DLPS
University of North Texas
Uppsala University

APPENDIX B: Survey Instrument - PREMIS Implementation Survey

This survey was developed by Preservation Metadata: Implementation Strategies (PREMIS), a working group sponsored by OCLC and RLG. The focus of PREMIS is on the practical aspects of implementing preservation metadata in digital preservation systems. One part of our charge is to develop a core set of preservation metadata with wide applicability within the digital preservation community. Another part is to examine alternative strategies for implementing preservation metadata. More information about PREMIS can be found at http://www.oclc.org/research/pmwg/.

This survey is being sent to organizations in the library, academic, museum, government, scientific and commercial sectors that are active or interested in digital preservation. It is intended to discover:

a) the different goals and characteristics of digital preservation repositories,

b) various administrative and technical models for digital preservation repositories, and

c) how preservation repositories are encoding, storing and managing their preservation metadata.

For the purpose of this survey, a "digital preservation repository" (also called "preservation repository") is any facility designed to store and safely preserve digital content for future use. "Preservation metadata" is metadata used by a preservation repository to carry out, document, and evaluate digital preservation processes.

This survey takes roughly an hour to fill out completely, although this will vary by institution. We understand this is a major investment of your time, and we thank you for supporting our work by responding. Survey results will be used by PREMIS in our analysis of implementation strategies, and will also be published in summarized form. Individual repositories and projects will not be identified without permission. We hope the results will benefit the entire community.

Please enter your answers directly on this Word document if possible, and return the document by email to **pcaplan@ufl.edu** by **January 16, 2004**. If you are not the appropriate person in your organization to complete this survey, please pass it along to the right individual.

Thanks very much,

Priscilla Caplan
Rebecca Guenther
co-chairs, PREMIS

Section 1: Contact information

1.1. If we need to follow-up on questions, who is the contact for this survey response (name, title, email)?

1.2. Does your institution have or plan to develop its own digital preservation repository?

__ Yes
__ No

1.3. Skip this question if the answer to 1.2 was Yes.

Does your institution use or plan to use an external digital preservation repository? What is the name of that repository? What institution is responsible for the repository?

Stop here. Thank you for responding to this survey.

1.4. Answer this and following questions only if the answer to 1.2 was Yes. (If it was No, answer 1.3 and stop there.)

What is the name of your digital preservation repository? What individual and/or unit within the institution has responsibility for running the repository?

Section 2: Business information

2.1. What is the mission of your preservation repository? Is the repository strictly dedicated to long-term preservation only or does it serve other goals too?

2.2. Who can deposit materials into your preservation repository? Please check all that apply:

__ general public
__ research community
__ my company or institution only
__ other companies or institutions
__ subscribers
__ other (explain)

2.3. What services does your preservation repository provide or plan to provide? Please check all that apply.

__ search and discovery
__ online, real-time access to service copies (restricted or unrestricted)
__ online, real-time access to archival copies (restricted or unrestricted)
__ secure storage of digital materials
__ data management of digital materials
__ storage and/or management of non-digital versions
__ preservation treatments (e.g. migration)
__ formal distribution of archival copies on request (real-time or batch)

__ reporting
__ billing
__ other (explain)

2.4. Does your preservation repository manage and track non-digital versions of digital materials as well?

2.5. How is the preservation repository funded? Please check as many boxes as appropriate apply. Note if funding model will change in the future.

__ grant funded external to institution
__ grant funded internal to institution
__ fee for service
__ part of organization's operational budget
__ other

2.6. What is the operational status of the preservation repository?

__ in the planning and organizational stage
__ in development (alpha, beta)
__ in production

If not in full production operation, when is it anticipated this will happen?

Section 3: Models and policies

3.1. What types of materials are accepted by this preservation repository? Explain in terms of all significant factors. Examples of significant factors might be file formats (e.g. .doc; .pdf; .jpg), authorship (e.g. government documents, faculty publications), ownership, date, publication status (e.g. published or working papers), object types (e.g. fixed format documents, web resources, applications, audio/video files), and subject area.

3.2. How are materials obtained by the preservation repository? Please check as many as apply and also explain.

__ harvested by repository
__ submitted to repository

3.3. What kinds of agreements does the repository have for obtaining materials? Please check as many as apply and also explain.

__ customers chose what to deposit
__ governmental deposit agreement
__ institutional archiving agreement
__ other legal mandate
__ other

3.4. Are there formal signed contracts or agreements with customers/depositors?

__ yes
__ no
__ not applicable

If yes, please provide an example if possible (send with survey response or attach below).

3.5. What are the policies or practices of the preservation repository regarding access to the materials? Check all that apply.

__ open access to all end users
__ access restricted to a particular community
__ access after a specified trigger event
__ on site access only
__ paid access
__ no online access
__ other

Describe any difference in treatment between preservation and access copies.

3.6. How is your preservation repository informed by the Open Archival Information Model (OAIS)? What features of the repository would you say conform to OAIS? What features, if any, do not?

3.7. In which ways do you find the OAIS reference model useful? Are there ways in which the model is unhelpful?

Section 4: Architecture, storage and preservation processes

4.1. What is the relationship between access and preservation copies of materials stored in your preservation repository? Please check all that apply:

__ access and preservation are served from the same copy
__ access copies are generated "on the fly" from preservation copies
__ access and preservation copies are stored in the repository
__ access and preservation copies are stored, but not in the same repository (explain)
__ there is no relationship because we don’t allow access ("dark archive")

4.2. Answer this section only if both access and preservation copies are stored. Do access copies get preservation treatment (e.g. migration)? Is the relationship between access and preservation copies maintained by the repository? If so how?

4.3. How are metadata and materials stored within the preservation repository? For example, one repository might zip metadata together with content files and store the zip file as a single entity. Another repository might store metadata in relational database tables and content files as individual entities in a file system. A third repository might use multiple models for different types of materials. Please describe all of the models that apply.

4.4. What preservation strategies are your preservation repository implementing now?

Please check all that apply.

__ restrictions on submissions (specified formats or quality)
__ making print or microform copies
__ bit-level preservation (secure storage, backing up, refreshing, etc.)
__ normalization (reformatting on ingest to more "preservable" formats)
__ migration (reformatting to more current version of the formats when the source format becomes obsolete)
__ migration on demand
__ emulation
__ other

Why did you chose these particular strategies?

4.5. What preservation strategies are your preservation repository planning to implement in the future? Please check all that apply.

__ restrictions on submission (specified formats or quality)
__ normalization (reformatting on ingest to more "preservable" formats)
__ migration (reformatting to more current version of the formats when the source format becomes obsolete)
__ migration on demand
__ emulation
__ other

Why did you chose these particular strategies?

4.6 What type of applications software is in use in your preservation repository? Please check all that apply. For commercially available or OpenSource software, please note what applications are used (e.g. "DSpace").

__ Commercially available software
__ OpenSource software
__ Locally developed software

Section 5: Metadata

5.1. What categories of metadata are (or will be) stored by and used by your preservation repository? Please check all that apply.

__ rights and permissions
__ provenance (document history)
__ technical metadata
__ administrative and management information
__ bibliographic/descriptive
__ structural metadata
__ other

5.2. If you are using or planning to use metadata elements from one or more published scheme, which schemes are you using? Please check all that apply.

__ AUDIOMD: Audio Technical Metadata Extension Schema
__ CEDARS
__ Creative Commons Metadata
__ METS
__ MIX or Z39.87
__ MPEG7
__ MPEG21
__ NEDLIB
__ National Library of Australia
__ National Library of New Zealand
__ OCLC Digital Archive Metadata
__ TEXTMD: Schema for technical metadata for text
__ Schema for rights declaration (METSRights.xsd)
__ VERS
__ VIDEOMD: Video Technical Metadata Extension Schema
__ other (please list)

5.3. Does your repository record information about these types of entities? Please check all that apply. Describe the sort of metadata that is (or will be) recorded about each of these entities, giving a few specific metadata elements as examples.

__ collection
__ logical object such as a book or photograph
__ non-digital source object
__ file
__ bitstream (a bitstream may be equivalent to a file, a subset of a file such as a binary object embedded in a PDF, or greater than a file such as a digital video stored in three parts)
__ metadata
__ other

5.4. How is metadata obtained (or expected to be obtained) by the preservation repository? For example, is it submitted by depositors, extracted automatically by the repository's computer programs, other? If different methods are used for different sets of metadata, please note all of them.

5.5. Do you require contributors to your preservation repository to provide metadata with their contributions? If so, what are your requirements? Please provide sample documentation if possible.

5.6. How is metadata stored and updated in your preservation repository? If multiple methods are used, please explain.

__ in a relational database
__ in an XML database
__ in an object-oriented database
__ in a proprietary database or format
__ in flat files
__ bundled with related content files

5.7. Some preservation repositories create or plan to create normalized or migrated versions of digital materials. If this applies to your repository, how will metadata for the original and new versions be handled? What information (if any) will be recorded only for the version created by the repository?

5.8. What metadata do you feel is most important to record for preservation purposes?

Section 6: If you want to add any comments to your survey response, please do so here.

APPENDIX C: Follow-up Interview Questions

(Note: Not all interviewees were asked all questions. Questions appropriate to each interviewee were selected from this master list.)

A few months ago you answered a survey on digital archiving for the PREMIS group. We have some follow-up questions to that survey and wondered if you could give us a few minutes to talk about them.

a.1) In your survey response you indicated that OAIS [whatever they said]. Could you elaborate on that?

a.2) What do you think it means to be OAIS-compliant?

a.3) Would you be interested in more information on practical implementations – a compilation of practices or Best Practices?

a.4) Do you think [libraries/museums] need another model, or another approach to the OAIS model?

b.1) Your response indicates that you rely to some degree on metadata supplied by the submitters to your repository. What metadata elements are these?

Do you have, or do you plan to have, any process in place for checking the accuracy of submitter-supplied metadata?

[Interviewer use these as prompts and check as appropriate:
___ Providing clearly articulated standards for submitters.
___ Offering assistance (e.g. consultation with staff members) to submitters needing help.
___ Including features such as drop-down menus for metadata capture.
___ Assigning staff to perform basic quality control over submitted metadata and correcting or returning problems to submitters.
___ Assigning staff to perform in-depth quality control and upgrading of submitted metadata, as needed.
___ Other (describe)]

b.2.) Which metadata elements are extracted automatically from submitted files? [ask for documentation if any]

b.3) When you obtain metadata both from the submitter and by automatic extraction, can this result in any duplication of information? If so, how do you/would you reconcile differences?

c.1) Your repository uses metadata to maintain the relationship between access and preservation copies of deposited materials. Can you describe that relationship in terms of how the copies are generated and stored? What metadata elements do you use to record these relationships? What are some typical values they might contain?

c.2) We are also interested in other types of relationships that you feel the repository needs to record:

Do you record intellectual relationships between objects, for example, objects that are part of the same collection, or objects that are different editions of other objects?

Do you record relationships between files, for example, different versions of the same file?

How do you record/express these relationships? (For example, through metadata, file names, directory structures, other)?

d.1) What factors influenced decisions about which types of metadata to store and use in your repository? Can you elaborate on how your metadata are stored [e.g., in relational tables, in XML, with or external to the files themselves]? What considerations drove those storage decisions?

e.1) Your preservation repository is using or plans to use migration as a preservation strategy. Migration may involve making multiple identical copies, making nearly identical versions in different file formats, or making versions in different file formats that lose some of the "look and feel", content or functionality of the original.

Is your agreement with depositors intended to give you the rights to carry out these actions?

[If the answer is yes:]

What words do you use to express these rights -- how do you interpret their meaning? [note the AHDS contract includes rights to "Electronically store, translate, copy or rearrange to ensure future preservation and accessibility"]

Is it important to stipulate that these rights are granted specifically for the purpose of preservation only?

[If the answer is no:] Why not?

e.2) Your preservation repository is using or plans to use emulation as a preservation strategy. Does the repository request specific permissions from the rightsholder to perform emulation?

[If the answer is yes:]

What words do you use to express these rights? Is it important to stipulate that these rights are granted specifically for the purpose of preservation only?

[If the answer is no:] Why not?

e.3) Are you relying on your digital content submitters holding copyright and/or authorization to grant intellectual property rights to protect you as the repository against possible violation of applicable copyright laws?

[If yes] Is the copyright ownership for digital objects submitted to the repository being documented by means of rights metadata in addition to any submission or depository agreement?

e.4) When the copyright holder is the repository itself (for internally digitized materials) or the greater organization to which it belongs, how is copyright ownership documented?

f.1) To what extent does your repository expect to document preservation related activities by describing and logging preservation “events”? For example, would you record it if you made a copy of a deposited file onto different media? What else would you record? What kinds of information would you record about these events? Do you distinguish event information from other types of digital provenance information?

[Thanks]