DIGITAL ELECTRONIC ARCHIVING:
THE STATE OF THE ART AND THE STATE OF THE PRACTICE

A Report Sponsored by:

International Council for Scientific and Technical Information
Information Policy Committee

And

CENDI

Prepared by:
Gail Hodge and Bonnie C. Carroll
Information International Associates, Inc.
Oak Ridge, TN

Final
April 26, 1999

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The exponential growth in the creation and dissemination of digital objects by authors, corporations, academicians, governments, and even librarians, archivists and museum curators, has emphasized the speed and ease of short-term dissemination with little regard for the long-term preservation of digital information. Digital information is inherently more fragile than traditional technologies such as paper or microfilm. It is more easily corrupted or altered, without recognition. Digital storage media have shorter life spans and require access technologies that are changing at an ever increasing pace. Because of these technological advances, the time frame in which we consider archiving becomes much shorter. Groups or individuals who did not previously consider themselves to be archivists are now being drawn into the role, either because of the infrastructure and intellectual property issues involved or because user expectations demand it.

This has raised the awareness of the issues surrounding digital archiving and preservation among information managers, librarians, publishers, and archivists. ICSTI, being a community which represents many of these information industries, has been involved in this issue for several years. Based on the most recent efforts by the ICSTI Electronic Publications Archive Working Group, this study was undertaken to provide information on the state-of-the-art and practice in digital electronic archiving.

Purpose, Scope and Methodology

In this project, "digital electronic archiving (DEA)" is defined as the long-term storage, preservation and access to information that was "born digital" (created and disseminated primarily in electronic form) or for which the digital version is considered to be the primary archive. [This does not include the digitization of material from another medium (such as digitization of paper or microfiche) unless the digital becomes primary.] Based on the analysis during this project, there is no common agreement on the definition of long-term preservation; the time frame is long enough to be concerned about changes in technology and changes in the user community. Depending on the particular technologies and subject disciplines involved, this time span may vary from 2-10 years.

The purpose of this study is to identify the state-of-the-art and practice related to DEA policies, models, and best practices, with an emphasis on the most "cutting edge" approaches. The study emphasizes those areas of most concern and interest to ICSTI members and those research areas previously identified by ICSTI as necessary to move the digital archiving discussion forward. Primary attention is given to operational and prototype projects involving scientific and technical information. The study is international in scope. It includes a variety of data types applicable to scientific and technical information, including data, text, images, audio, video and multimedia, and a variety of object types, such as electronic journals and monographs, satellite imagery, biological sequence data, and patents.

The study methodology involved an initial survey of the ICSTI and CENDI members (see Appendix A-1 for a full copy), as well as a literature review and contacts with experts, to identify the most "cutting edge" projects. The highlighted projects cover six countries (U.S. (9), UK (2), Canada (1), Australia (1), Sweden (1) and Finland (1)). Four organizations are considered to be international in scope, because their funding sources and scope are not bound to a particular country. The projects come from a number of sectors including government scientific and technical programs, national archives, national libraries, publishers, and research institutes. Information about other projects is included where applicable.

After the initial questionnaire, follow-on discussion questions (see Appendix A-2) were developed and aimed at identifying emerging models for the relationship between the various entities in the information chain (users, intermediaries, primary publishers, secondary publishers, online vendors, and others) as they relate to archiving; the metadata information that is being gathered; how the archive will be maintained and accessed; an estimate of the costs to be incurred for start-up and maintenance; and outstanding issues and possible best practices. While technologies for storage and retrieval may be mentioned in the report, technology is of secondary interest to the understanding of policy and practice.

General State of the Art/Practice

The issue of archiving digital objects brings together several normally diverse communities -- archivists, records managers, librarians, data center managers, and data producers. There is so much activity among various groups that it is difficult to encapsulate the general state of DEA. However, there are a few general models that can be highlighted as emerging. The models have genesis in one of the diverse communities, but may have applicability to others.

It is noteworthy that many of the major projects in digital archiving are of a cultural or historic nature. While the emphasis in this study has been on scientific and technical projects, the humanities-related projects have provided the basis for much of the current thinking in this area. They have been used peripherally in this study for what they offer to science and technology, or for the scientific and technical information components that many have.

Identified Organizational Models

The highlighted projects were analyzed for commonalities that would identify organizational models for DEA. The approach taken is an organizational one, loosely based on the previous work sponsored by the Arts and Humanities Data Service (AHDS) (Beagrie and Greenstein, 1999). Four major organizational models -- Data Centers, Institutional Archives, Third Party Repositories, and Legal Depositories -- were identified. An additional conceptual model for interoperable archives is also described. These models are based on differences in the information flow, the management of the life cycle functions of the archive (creation, management, preservation, and access), responsibility and ownership of the data, and the economic model.

The most mature archival model is that of the data center. Three subcategories of data centers were identified based on the degree of homogeneity and centralization. Centralized data centers, such as the National Digital Archive for Datasets (UK), have numerous contributors, but a central repository and administration. This model allows for easier integration of the data and more consistent adherence to standards. However, there may be little backup for the central repository, particularly if funding is cut. It is also difficult to include new data producers with varying data models, standards and primary audiences. Federated data centers, like the NASA Distributed Active Archive Centers (DAACs) operate in a distributed, but closely guarded environment with common standards and practices, and a single user interface. There is redundancy in the federation's ability to respond to user needs. With looser standards, more partners may be involved more easily. Cooperative data centers do not currently exist, but there is a prototype under development between the San Diego Supercomputer Center, the U.S. National Center for Environmental Analysis and Synthesis, and the Long Term Ecological Research Network. The aim is minimal metadata and system standards, acknowledging the diversity of data types, models and structures in ecological science.

On the whole, the data centers are also the simplest organizational model. The intellectual property rights are generally clear, because the owner is the funding agency. The economic model allows free access by the funding agency and, since many of these are government sponsored or internationally developed data banks, the public also has free access. Additional charges may be levied for extraordinary services or for access for commercial purposes. However, it is unclear how well the practices of these data centers, which have large volumes of relatively simple data, would migrate to other communities and object types.

Institutional archives are generally a department or branch of an institution that collects and preserves the intellectual capital for that institution. These institutions can include publishers, data producers, societies, cultural organizations, government agencies, academic institutions and industries of various types. Institutional archives generally have some level of ownership of the information. Often access is limited to members of the organization, to subscribers, or to partners in a particular project or venture. Many corporations and institutions archive only what is required by regulation, fearing legal ramifications if certain information is retained. However, there are organizations such as pharmaceutical, chemical and petroleum companies, where internal scientific and technical information is critical to the perpetuation of research and development. Institutional archives may also increase as the knowledge management technologies connected to intranets reach a wider market.

Third party repositories are the third model. They tend to derive from the journal publishing industry, rather than government data centers or institutional records needs. They can be divided into two types: Publication Service Providers and Repository Management Agents.

Publication Service Providers serve other roles in the information chain. In addition to their primary role as vendors, electronic publishers, or jobbers, they may also provide digital archiving as a service to their clients, which are primarily learned societies and publishers. This is the most

complex organizational model for archiving, because there are numerous roles being played by the participants. Often the economic model for the archiving is not clear, because it is bundled with the other services that the publication service is providing. Examples of Publication Service Providers who also provide archiving services include ingenta, Ltd. and HighWire Press.

Repository Management Agents are an emerging model in DEA. These organizations act as trusted third party repositories, but do not serve any other function in the value-added chain. They provide a safety net by continuing to provide access to the digital object should the publisher or producer of the object determine that it can no longer archive the material or if it goes out of business. Examples include JSTOR and OCLC's Electronic Journals Project. Both projects have substantial numbers of journals available. The majority of JSTOR's current titles are in the humanities and social sciences. However, they have recently begun a project on a Science Cluster, which will include AAAS's *Science* and the publications of the National Academies of Science (US). In both cases, the charges are borne by the user or library. JSTOR's pricing model is based on a yearly subscription to the JSTOR service, with rates differing by size of institution. OCLC's model is based on the library's subscription to the electronic journal directly through the publisher, through a jobber, or in some cases through OCLC. The agreement requires linked access to the publisher's archive or deposit of a digital copy with OCLC. OCLC is currently working on the long-term business and pricing model for this service.

The fourth model is that of the Legal Depository. There are generally two types of legal depositories: national depositories and national libraries. The national depository (or archive) has tended to document the business of government, which includes administrative documents. The national libraries are generally charged with maintaining the culture, history and intellectual output of the country by collecting what is published within that country. Both national libraries and national depositories have sought to handle digital material. As part of digital government initiatives, archives such as the UK Public Records Office and the U.S. National Archive and Records Administration have extensive electronic projects. In the UK, the PRO has separated the responsibility for archiving digital datasets from the archiving of digital office records.

Some national libraries have sought to extend their mandate to digital information. In many cases, they are doing this without the benefit of legislation. The PANDORA Project of the National Library of Australia has the most extensive guidelines for the selection of Web-based "Australiana". The National Library of Canada's Electronic Collection incorporates electronic books and journals published in Canada in its regular workflow, based on the results of the Electronic Publications Preservation Project pilot study. The National Library of Sweden is using robots to harvest all relevant domain names and Web servers, archiving the content without review. Projects are also underway at the National Library of Finland. The Networked European Depository Libraries (NEDLIB) project is funded by the European Union to investigate the procedures, standards and infrastructure needed to support a multinational library network for digital archiving.

Though not an operational model, the interoperable archive model described in the recently

drafted *Reference Model for an Open Archival Information System (OAIS)* (Consultative Committee for Space Data Systems, 1999) provides insight into the future of a hierarchy of archival organizations and heterogeneous archives, and is worth evaluation in this context. This reference model provides terms of reference, conceptual data models, and functional models for open archives that can interoperate. The models are based on packets of information, including the data object itself, descriptive metadata, representational information which helps to interpret the bits in the data object (e.g., the ASCII table), and specific information needed for preserving the object. Based on the exchange of these packets, and the standardization and crosswalks among the metadata formats used to present the information, objects can move from one archive to another, and archives can be searched simultaneously. Many experts, including the CEDARS project in the UK, are investigating whether this data-centered model could be generalized across other data types.

Life Cycle Managers and Their Roles

The results of the study were also analyzed for the changes in the roles of the traditional players in the information dissemination chain. The roles analyzed include creator (author), publisher, secondary publisher, library and consortia, funding source, and user.

The analysis found that creators and users are not very involved in the digital archiving process. However, this is changing as organizations are requiring metadata creation with digital objects, and as software is developed to make the creation of such metadata (and even its automatic extraction) easier.

Publishers are involved in digital archiving in a number of ways. The most vocal are the learned society publishers who consider this to be part of the mission for their discipline or organization. However, the economics and long-term viability of such preservation (as the content of the system grows) is unclear.

Few secondary publishers have expressed an interest in digital archiving according to an informal study conducted by the National Federation of Abstracting and Information Services. However, many of these services have a long history of migrating and maintaining archival collections of bibliographic records in a discipline. Third party repositories, particularly OCLC, and national libraries (the National Library of Australia) have designed systems to take advantage of the bibliographic records as the catalog record that provides access to the full archival object.

Libraries, particularly consortia, have been instrumental in raising digital electronic archiving issues. As they seek to provide access to electronic journals, which no longer provide a consistent physical copy that can be owned and preserved, libraries have developed guidelines for license agreements which include statements regarding digital electronic archiving. Licenses generally provide for a trusted third party or the library itself to receive and archive an electronic copy immediately or when it is no longer available from the publisher.

Funding is a key driver in the evolution of archive models. Funding is provided by government organizations, national and international science initiatives, private foundations, research institutes, and museums. Funding organizations in many quarters have espoused the need for archiving digital information. Unfortunately, in many cases, particularly at the government level, there have been mandates without supportive funding. In many cases, guidelines have been developed, but they are not detailed enough to provide real guidance on issues of long-term preservation, media migration, and planning for the related costs in program and project budgets.

Best Practices

The evaluation of the research results was organized by again looking at the best practices by the information life cycle for archiving material across the various models. The life cycle functions are creation, acquisition/collection development, cataloging and identification, storage, preservation and access.

Practices used when a digital object is created ultimately impact the ease with which the object can be digitally archived and preserved. The preservation and archiving process is made more efficient when attention is paid to issues of consistency, format, and standardization in the very beginning of the information life cycle. Institutions are beginning to require a more limited number of formats for some objects created under their auspices.

All groups involved acknowledge that creation of good metadata at the source of data creation is where the long-term archiving and preservation must start. As standards groups and vendors incorporate Extensible Mark-up Language (XML) and RDF (Resource Description Framework) architectures in their word processing and database products, creating metadata when the digital object is created will be more efficient and more rapidly adopted. However, work remains to identify the specific metadata elements needed for long-term preservation, particularly for non-textual data types likes images, video and multimedia. Others in the information creation chain for formal materials, e.g., publishers, funding sources, learned societies, etc. can play a large part in promoting such attention on the part of creators and the development of relevant preservation standards.

Cataloging and identification issues are often interrelated with decisions about what to archive and how long it will likely be retained. The metadata to be collected, and the degree to which a standard will be used, depends on the type of organization doing the archiving, the resources available, the type of material to be used, and the requirements of funding organizations. The most common formats are MARC and Dublin Core. Only the traditional publishers appear to be using the Digital Object Identifier. Other stakeholders have developed their own identification schemes.

The national libraries are taking the lead in the development of guidelines related to the acquisition and collection of digital objects in archives. The PANDORA project has extensive guidelines for a variety of Web-based (primarily textual) material, including ephemera. Issues

addressed in the guidelines include determining what should be archived, determining the extent or the boundaries of the digital work, and archiving related links.

Storage issues center around hardware/software migration. New releases of software can be expected every 2-3 years. Migration to new media and hardware occur less frequently, but can be expected at least once every 10 years. The general response from those queried about these issues is that they have no firm plans for migration, but will plan to stay up to date with current technologies by migrating the content to each new technology. The issues of cost have the biggest concern here, and there is now a sense of having to deal with it as best we can as the technologies change. All the respondents followed industry best practices related to refreshing the media, back-up, recovery and remote storage for disaster recovery.

Preservation is the aspect of archival management that preserves the content as well as the look and feel of the digital object. In cases where the archiving is taking place while changes or updates may still be occurring to the object, such as with datasets or electronic journals, attention is being given to refreshing the site contents. The National Library of Australia allocates a gathering schedule to each "publication" in its automatic harvesting program. Obviously, the burden of refreshing the contents increases as the number of sources stored in the archive increases.

Most organizations lack formal retention policies, because they are relatively new to digital information and storage costs continue to decrease at a faster rate than the increase in the size of most archives. The most common answer is that the organization will archive "everything" for "all time". Other than legal depositories, there is little recognition of the need for more definitive policies in the future based on the value of the information to potential users, the resources available on the part of the archiving organization, and the desires of the funding agency. Those who recognize the need for such policies also acknowledge that we do not have a crystal ball, and, therefore, it is difficult to determine precisely what will be of value in the future. When the burden gets too great, particularly for commercial institutions, it may be necessary for public institutions to intervene and provide a backup archiving service for objects that are no longer of sufficient commercial value to warrant inclusion in the commercial organization's archive.

Preservation has also involved the decision of whether to transform the incoming information into a new, more standardized format, or to retain the native format. While the answer to this depends to some extent on the user community being served by the archive, and the degree to which the transformed format matches the native format, there appears to be a tendency to transform to the newest related format, for example from the current version of TIFF format to the next. However, in some cases where legal responsibilities intercede, the original is always retained, along with the transformed format for access.

Regardless of the decision about transformation versus native format, preserving the "look and feel" of the object remains an issue. If the digital information is transformed, the question is how much does this impact the "look and feel"? If the information is retained in native format, how

will the "look and feel" be provided when the technology changes in the future. Migration is the most common answer to this issue, realizing that the "look and feel" may not always be retained. An alternative is an emulation strategy. Emulation involves reconstructing the behavior of the hardware and software in the future environment in order to recreate the "look and feel" of the original digital object in its old environment. This will involve cooperation on the part of hardware and software vendors to provide access (or perhaps restricted registries) to proprietary information about the hardware and software. However, to-date there have been no large-scale pilot projects that would indicate that the emulation approach is practical or scalable.

Finally, the life cycle of archived material requires access or the ability to reuse the information. Currently, all projects reviewed have or are planning Web-based interfaces to their archives. Additional interfaces are available for certain specialized information, such as the datasets available from the data centers. However, digital archivists are looking beyond the Web to another as yet unknown interface, and they consider the interface to be another technology that can change rapidly.

Depending on the intellectual property and licensing issues, the access to the objects may be restricted. Archives that store copyrighted materials, proprietary information, or restricted government information must also deal with security and authentication issues. Processes being investigated or put into place may include digital signatures and certificates, in addition to the more traditional IP address and user name/password log on procedures. The ability to download and reuse the information also differs depending on the archive, the license agreements with the rights holders, the type of user and his relationship to the archive or rights holder, and the amount and type of material being downloaded. Because of the ease with which digital material can be altered, either knowingly or unknowingly, mechanisms such as watermarks or encryption are viewed as key tools in the process of digital preservation.

Best practices are also beginning to emerge for different format and object types. Image archives are particularly concerned with the type of metadata information needed for preservation and access to these images, including changes in resolution and compression techniques. The Research Library Group, the Digital Library Federation, and the U.S. National Information Standards Organization, partnered with a variety of European organizations, are involved in developing such guidelines and metadata elements which will be available for review in the next few months.

All the issues related to the various data types, and more, are bundled into the issues surrounding the archiving of multimedia works. Since efficient archiving, access, reuse and preservation differ based on data type, multimedia, which combines various data types, cannot be dealt with by a single approach. In addition to the archiving of a series of objects that make up the multimedia object, it is important to be able to bring the collective multimedia object back together again. Projects in this area are underway within the US Department of Defense and the US National Library of Medicine. A standard file format for multimedia is being developed by Microsoft.

Costs/Resources

Although cost is recognized as a basic driver in DEA, it was also the most difficult aspect on which to gather information. In some cases, a lack of response was because of the proprietary nature of this information. However, in most cases, the respondents indicated that they just didn't know how much the archive was costing or would cost in the future. For publishers and producers, the cost of archiving is still tied up in the cost of manufacture. This is also true of publications services where the archiving is considered an added benefit to the publishers who are served. Until several large archives have gone through at least one or two migrations or emulation developments, it will not be possible to separate the cost for the archives from the cost of doing business.

Anecdotal information is available from several national library or institutional projects that are archiving Web sites, electronic journals and other digital publications from the Internet. However, the information is generally presented in terms of the number of full or part-time staff being devoted to the effort at this time, with no indication of hardware/software or other infrastructure costs now or in the future.

In addition to questions of start-up and ongoing operation, there is a serious issue of the long term financial commitment to archives. Increasing recognition by scientific authors and funding sources is key to the success and sustainability of an archive. Several experts interviewed suggested that an endowment model might be needed. This would set aside a portion of the payment for the use (whether storage or access) of the archive for its perpetual care.

Conclusions

Based on the analysis of the organizational models, the changing roles of traditional stakeholders, and best practices in digital life cycle management, general conclusions can be made in the areas of most interest to ICSTI. These include policies, organizational models, and economic models.

The policy issues of major concern seem to be the intellectual property issues, and with them the related security and authentication concerns. To greater or lesser degrees, all stakeholders in the archiving and preservation chain are concerned about intellectual property. For many of the data centers, the issue is put in public versus commercial use terms, and is reflected in the types of access and services provided and the charges placed on them. For publishers and producers, intellectual property concerns are reflected in the kinds of business arrangements used to promote their archives. Intellectual property concerns have led some organizations to consider institutional archives, where the information remains under their control. Others, lacking the resources to do this, but still concerned about their intellectual assets, are contracting with publication services or trusted third-party repositories. Part of these contracts requires security and authentication on the part of the archive, as well as specific procedures for granting and continuing access. Libraries, consortia and users are increasingly attuned to intellectual property issues, and their concerns for fair use in a digital environment are often reflected in the license agreements that are signed.

Five organizational models for digital archiving have been identified. Aggregation on the part of repositories, publication services and legal archives is likely to continue as stakeholders struggle with how to make the information accessible with common interfaces, in the midst of cost and intellectual property concerns. Based on the numbers and types of organizations involved, the need to integrate across format and object types in the sciences, increased emphasis on multimedia, and ever-changing technologies, the organizational model for archives in the foreseeable future appears to be a loose network of archives covering special disciplines, geographic areas, or object types. Using network technologies and interoperable standards, the future model will likely be a network of disparate but interoperable archives. Individual communities are likely to develop standards and common practices. Interoperability in a heterogeneous environment is likely to be required. The Open Archive Information System (OAIS) reference model, described earlier, appears poised to promote this interoperability beyond the realm of data-centered archives.

Similarly, it is likely that there will be a variety of economic models for digital archiving. This will impact not only the way the archives are managed and who manages them, but the value (and the cost) involved in retaining older materials. Some archives will be commercially viable, others will not. Some will need to charge for services, while others will not. When archives are governmentally appropriated, there is increasing recognition of a long term maintenance commitment, but there does not seem to yet be sufficient definitive action and funding to support this recognition.

With a large number of models and increased interest in the future of digital information, many stakeholders are getting into the archiving business. There are many organizations that appear to consider this a reasonable avenue for business growth. With the large infrastructure and varying skills needed to perform digital archiving satisfactorily, we may be seeing the rise of a new industry. Smaller publishers in particular may continue to look for avenues by which they can contribute to one or more archives, without undertaking the infrastructure development themselves.

Multiple economic and organization models are likely to persist in the DEA environment. As the report of the ICSTI Electronic Publications Archive Working Group suggested, a hierarchy of archiving organizations may be needed to overcome the economic and intellectual property issues that continue to abound in the digital environment.

It appears the discipline specific, as well as national and global archives, will be built incrementally on the basis of pilot projects that lead the way and evolve into a complex network of content infrastructure. The issue has been recognized and the bandwagon is growing. In summarizing best practice areas, we see building blocks for future developments. The trick will be the coordination of these archives to reduce the expense of unnecessary redundancy, to tie the system together in an integrated fashion for the user, to ensure long-term funding for these archives, and to mechanisms to protect the rights of both copyright holders and users.

Recommended Next Steps

Based on the survey and analysis conducted during this project, the following actions are recommended for consideration.

ICSTI

1. **Many models are evolving and taking hold. Each stakeholder will be affected and the activities should be monitored for more specific and ongoing relevance to ICSTI member groups:**

- **Hold discussions on impacts of the various models (both organizational and economic) for classes of ICSTI members. Monitor projects selected by members to be models for their part of the industry, and provide opportunities for interaction between these projects and appropriate communities within ICSTI.**

Projects that include the specific stakeholder group or the portion of the information life cycle function in which a particular organization is interested should be monitored with specific reports back to ICSTI members interested in these particular areas. In addition to project monitoring, opportunities should be provided for interaction between the project managers of the selected projects and ICSTI members. The next annual meeting, or a special meeting cosponsored with ICSU, UNESCO or some other organization, would provide a forum for the discussion of these specific projects. It might also be valuable to hold the session concurrent with a major meeting where these projects might already be represented.

- **Interpret the draft Open Archive Information System (OAIS) Reference Model for the ICSTI Communities**

Since heterogeneity and a complex network seem to be evolving, the OAIS Reference Model is one worth further group exploration. It stands as a possible framework for data interchange needed across the various functions of an archive (regardless of the players involved), and across archives. However, the current reference model is still very data-centered. ICSTI should convene a small group or groups of stakeholders to interpret the reference model for the different communities -- primary publishers, secondary publishers, and libraries. During this process it should be possible to determine if the reference model has utility for a variety of stakeholders and a variety of data types. The CEDARS project in the UK has expressed an interest in working together with ICSTI on this review. This follow-on project should be done in the context of the ISO review of the draft reference model and should consider interoperability, standards, common practices and economic models that will have to coexist. The benefit to ISO and the Consultative Committee on Space Data Systems is that they will obtain a review by an expert community, outside the data community. The benefit to ICSTI is that it may find a model that can be used across its members and to inform the community at large.

- **Develop a Digital Electronic Archive Registry Emphasizing Digital Publications**

The Electronic Archive Registry, recommended by the ICSTI Electronic Publications Archive Working Group, may act as a transitional mechanism between the current distributed, unintegrated archiving projects for electronic publications and the fully networked environment envisioned by the OAIS. The Working Group envisioned this registry as a finding aid for the location of where, by whom, in what format, and what parts of a publication are electronically archived. The data elements required for such a registry and the procedures whereby the registry is created, maintained and accessed must be developed. The Working Group suggested that the registry could be added to the ISSN system. The concept should also consider the work of other groups such as the Digital Object Identifier (DOI) Foundation and the national libraries/bibliographies.

- **Monitor and report on the key projects related to the cost and organizational issues of digital archiving**

This review has identified that there are still significant unanswered cost and economic questions related to long-term digital archiving. Some of these questions are related to the speed of technological change, while others are institutional. However, there are several significant projects under way that have been briefly identified in this report. They should continue to be monitored and progress on them reported to the ICSTI community. Recommendations for projects to be monitored include NEDLIB, the objective of which is the networking of depository libraries and the development of digital depository format standards for publishers; CEDARS, which is looking at the networking of UK archives; and Cornell University's Digital Library 2-Initiative which will address cost and organizational issues. Relationships should be established with these projects in order to learn about their progress and be able to report on the outcomes to the ICSTI listserv.

2. As appropriate, work at individual organization levels to promote digital archiving practices:

- **Recommend to ICSTI organizations that digital standards for metadata and object identification that are under consideration be reviewed with a particular eye to their ability to support long term preservation and access.**

In particular, work to ensure that the concept of archives and preservation is developed and used within existing and forming standards for metadata and identifier.

- **Provide testbed material for projects when possible.**

A significant way for ICSTI members to become involved and to learn more about the challenges and best practices in this area is to provide material for digital archiving testbeds. This is already being done by Elsevier, Kluwer and Springer in the NEDLIB project. There may be similar

opportunities with other projects, including CEDARS and the Cornell University DLI-2 projects.

- **Promote multilateral projects, to promote the development of best industry practices in digital archiving**

Promote round-table sessions at a follow-on ICSTI meeting that would bring together ICSTI members working on similar issues related to digital archiving so that resources, lessons learned, and pilot projects could be shared. Of particular importance would be discussions and pilot projects related to business models for digital archiving and intellectual property issues (particularly between national libraries and publishers).

Both ICSTI and CENDI

1. **Make ICSTI/CENDI's interest in this area known so the organizations stay involved with the forefront of activities and continue to keep the debate visible with customers, suppliers, and funding sources.**

- **Present a paper at the World Science Conference**

As suggested by the ICSTI Executive Board and planned in the proposal, the results of this study will be presented by Dr. David Russon at the World Science Conference in July 1999.

- **Develop a Statement of Concern regarding digital electronic archiving**

As many survey participants mentioned, the current projects in digital electronic archiving are often being done without adequate commitment and funding. There is concern that funding will not be sustained, and is not consistent with mandates to collect and preserve electronic information. As suggested by the ICSTI Working Group, ICSTI and CENDI should produce a Statement of Concern, either jointly or consecutively, that raises the issues of electronic archiving and continued preservation and access to these archives with stakeholders, policy makers and funding sources. Many of the stakeholder groups are represented by members of ICSTI and CENDI, and therefore, it should be in a unique position to "work through" this difficult task. As the ICSTI Digital Electronic Working Group indicated in its report, the statement should not only identify the need for and benefits to be gained by electronic archiving and continuing access, but it should identify guidelines for what constitutes an electronic archive and sufficient access. It should emphasize the need to support verbal commitments to digital archiving with proper programming and funding. The Statement of Concern should also identify further activities in which ICSTI and others can participate to ensure that the statement is put into action.

- **Publish an article on the results of the ICSTI/CENDI study**

While the report to the World Science Conference will provide some level of visibility for the efforts of ICSTI and CENDI as well as for the next steps necessary to move digital archiving

forward, this will not reach all stakeholder audiences. It is suggested that an article be prepared from the study and published in a relevant journal. The investigators have already been approached by the editor of the *Journal of Electronic Publishing* for such an article.

- **Develop a topical area on either the open ICSTI or CENDI Web site that highlights digital electronic archiving.** (This could also be done as a joint effort.)

The topic of archiving was highlighted in the report from the June 1997 meeting and in a subsequent issue of the *ICSTI Forum*. Those documents, a summary of this report and other possible information gleaned from ICSTI members should be included as a special theme on the Web site. (There are many good sites that already address this issue, and there is no need to replicate them. However, links from a specific ICSTI or CENDI page to these other sites may be of value to ICSTI and CENDI members and others interested in this subject.) CENDI could consider highlighting this area as a special adjunct to the broader STI Manager part of its Web site.

This survey has emphasized that DEA issues require collaboration and coordination among a variety of stakeholders. There are numerous projects underway at many levels. The ICSTI and CENDI members can benefit from staying informed of ongoing activities. They also have experience and practical needs that can help to inform and move the state of DEA implementation forward.

Introduction

> **As we move into the electronic era of digital objects it is important to know that there are new barbarians at the gate and that we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever. We are, to my mind, living in the midst of digital Dark Ages; consequently, much as monks of times past, it falls to librarians and archivists to hold to the tradition which reveres history and the published heritage of our times. - Terry Kuny, XIST/Consultant, National Library of Canada**

Several information scientists, historians, and philosophers have begun to speak of our time as the digital dark ages. Similar to the period before the printing press, where a small group (monks and scribes) preserved what they could of the cultural heritage, which was lost through the imperfections of the oral tradition, we are facing similar losses of our heritage, not just cultural and historical, but scientific and technical. However, unlike the 15th Century where the savior was technology in the form of the printing press, the losses of the 20th century and those that will persist into the 21st are being caused by a technology, used without consideration for the future.

The technology or series of technologies that have created such fragile information are personal computers, electronic publishing software, and, most of all the Internet. The exponential growth of the use of these technologies by authors, corporations, academicians, governments, and even librarians, archivists and museum curators, has emphasized the speed and ease of short-term dissemination with little regard for the long-term preservation of digital information.

There are several aspects of digital information that make its archiving different from paper. Digital information itself is considered by many to be more fragile than traditional technologies such as paper or microform. While there are problems like acid paper that affect paper archives, and changes in microfilm techniques and reader equipment that impact the use of microforms, there are new and different challenges when the material exists only in electronic form. It is more easily corrupted or altered. Digital media, such as CD-ROMs, may have a shorter than expected life-span. Digital information is more susceptible to changes in the technologies of access and retrieval. In some cases, the information is so closely linked to the software or other technology that it cannot be used outside these proprietary environments. (Kuny, 1998).

Because of these technological advances, the time frame in which we consider archiving becomes much smaller. The time between manufacture and archiving is shrinking. Groups or people who did not previously consider themselves to be archivists are now being drawn into the role, either because of the infrastructure and intellectual property issues involved or because the expectations of users groups demand it. "...we have rarely had a preservation imperative arise so quickly after original manufacture, especially on such a large scale, as we do with digital materials. Relative to the other materials that tend to find their way into museums, archives or libraries, we will not have the benefit of a tradition of care and maintenance that will guide our actions when it comes to digital works." (Messier, 1998)

Because of the fragility of digital media, the lack of a tradition of stewardship, and the speed with which electronic publishing has grown worldwide, archivists, publishers, and librarians have become increasingly concerned about the archiving and preservation of digital information. Since many of these communities are represented within ICSTI, it is not surprising that ICSTI has been involved in this issue for several years. [1] At the December 1998 meeting in London, the ICSTI Board approved a study of the state-of-the-art and practice of digital electronic archiving, as a follow-on to a paper presented at the meeting by the Electronic Publications Archive Working Group (ICSTI, 1998). Based on common interest in this topic, CENDI, a U.S. interagency working group of scientific and technical information managers in the federal government, approved co-sponsorship of this study in February 1999.

---

[1] After the 1996 UNESCO/ICSU Meeting on Electronic Publishing, ICSTI was approached by ICSU to investigate the topic of electronic/digital archiving. In response, ICSTI made this the topic of the 1997 Annual Meeting in Philadelphia. The technical sessions were centered around this topic, and presented many projects that were going on world-wide by various groups within the information community – learned societies, authors, commercial publishers, A&I services, librarians, etc.

By holding this session, ICSTI identified several areas where additional research was needed. A working group was proposed to continue research, to gather information and to forward recommendations to appropriate bodies. In 1998, the Electronic Publications Archive Working Group was formed. The Working Group met during 1998, and the report was presented at the Winter Meeting. The group addressed the research areas identified in the June 1997 symposium and identified several possibilities for ICSTI involvement. "Possibilities discussed by the Working Group included helping to spread the word through a Statement of Concern; gathering information on what plans publishers, libraries, etc. have made to date regarding electronic archives; planning and helping to start a "registry" of publications with information on if, how and where they are archived; and finally to join with others in a session on the subject at a major scientific meeting in 1999."

The Working Group acknowledged that other groups have continued to work on digital electronic archiving issues since the 1997 symposium. They felt it was important to put any additional ICSTI activities in the most current context possible. The report suggests that an effort be made to gather information about what other groups, both inside and outside ICSTI, are doing or planning to do related to electronic archiving. This information would provide input to other possible follow-on tasks suggested by the report – the registry and the Statement of Concern, and would also contribute to the working knowledge of ICSTI members and others concerned with the transition from print to electronic publishing.

A proposal submitted by Information International Associates, Inc. of Oak Ridge, Tennessee (IIa) to survey the state of the art and recommend areas of further involvement by ICSTI and its member organizations was originally presented at the ICSTI Winter Meeting in London, and was refined based on discussions with the ICSTI Digital Archiving Advisory Group. It addressed two of the areas of focus for ICSTI's next activities with regard to digital archives: gathering information on operational digital archives and identifying their characteristics and experiences.

At the same time ICSTI was considering the proposal to undertake this study, a U.S. based group of federal scientific and technical information managers known as CENDI was considering similar issues. In February, CENDI joined ICSTI in co-sponsorship of this study.

What is Digital Electronic Archiving?

[An archive] consists of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for one or more designated communities. - Reference Model for an Open Archival Information System, ISO Consultative Committee for Space Data Systems

During the course of the survey, it became evident that the term "archiving" itself has taken on many overtones that color people's perceptions when this term is used. In some cases collections of material were called archives regardless of whether or not the organization had considered long term storage and preservation. The simple use of the noun "archive" does not result in an organization being attentive to the archiving of the collection, or taking an archivist role. NASA has gone as far as calling its centers "active archives" which provide the underlying notion that the data was collected for active use and has the function of maintaining the information for this use – hence an "archive."

"Digital archiving" or "digital preservation", terms which tend to be used synonymously, refer to the long-term storage, preservation and access to digital information. In this project, "digital electronic archiving (DEA)" is used to narrow the scope to focus on information that was "born digital" (created and disseminated primarily in electronic form) as opposed to projects that digitize material from another medium (such as digitization of paper or microfiche). Based on the analysis during this project, long-term preservation has no specific time limit; it is long enough to be concerned about changes in technology and changes in the user community. (Depending on the environment, this may be only a time span of 2-10 years.)

Purpose and Scope of the Study

The problem exists at every level, from small business to great archival institutes to the ordinary household. You can't simply cram all this information in a box and stick it in the attic, because the attic is already jammed, as are the basement and all the closets. - Joel Achenbach, "The Too-Much-Information Age", Washington Post, March 12, 1999

Given the breadth of the digital archiving challenge, it was necessary to narrow the scope of the study to emphasize those areas of most concern and interest to ICSTI members and those research areas previously identified as of value to moving the digital archiving discussion forward. Therefore, the purpose of this study is to identify the state-of-the-art and practice related to DEA **policies, models, and best practices, with an emphasis on "cutting edge" approaches**. For the purposes of this study DEA is defined as the long-term preservation of information published ("born digital") or communicated initially in electronic form (and perhaps in print as well). It does not include projects that simply convert legacy print information into electronic form for preservation and archiving. However, DEA may apply if the resulting electronic version is considered to be the primary or sole archive.

The study focuses on **scientific and technical information**. However, it acknowledges that

there are some projects of mixed origin, particularly within the depository and national libraries, and that work in the humanities and social sciences can be used to inform this discussion.

Because the challenge of DEA extends beyond anyone country's borders, the study is **international** in scope. The call for projects was sent out worldwide.

The study includes a wide **variety of data types** applicable to scientific and technical information, including numeric data, text images, audio, video and multimedia. It also includes a **variety of document types**, including electronic journals, monographs such as technical reports, ecological and environmental datasets, satellite imagery, biological sequence data and patents.

Projects were selected based on the use of emerging models for the relationship between the various entities in the information chain (users, intermediaries, primary publishers, secondary publishers, online vendors, and others) as they relate to archiving; the metadata information that is being gathered; how the archive will be maintained and accessed; an estimate of the costs to be incurred for start-up and maintenance; and outstanding issues and possible best practices. While technologies for storage and retrieval may be mentioned in the report, technology is of secondary interest to the understanding of policy and practice.

The primary audience for this report is the ICSTI and CENDI memberships, with a secondary focus on a presentation to be given to the World Scientific Conference. It is expected that the results will also be of interest to a broader audience and that the findings should be usable by ICSTI to determine what role it might play in further efforts regarding digital electronic archiving. The report will also be shared with those who participated in the study.

Data Collection and Analysis

This project involved extensive data collection including a review of the literature, contacts with experts and two questionnaires. The methodology is described below. Because of the extensive information available related to DEA, it was necessary to highlight several key projects. The selection of these projects is also described.

Methodology

The data collection occurred in two phases. The first phase sought to cast a broad net and identify projects that might be relevant to DEA. The second phase included more detailed follow-up on the projects that looked most promising. In each phase, information was gathered through surveys of the ICSTI and CENDI members, literature searches, and contacts with experts.

In the first phase, an initial survey (see Appendix A-1) was sent via listserv to ICSTI and CENDI members. The survey was intended to identify possible DEA projects both within the member organizations and those known to members within their subject disciplines or geographic regions. Of the 55 ICSTI and CENDI members, 18 responded. The survey was also sent to other listservs

including NFAIS, Dig-Lib, ASIS-L, ASLIB, IFLA, and ARL, and to key members of the Society for Scholarly Publishing which does not have a listserv.

An initial literature search was also conducted of both traditional published literature and Web resources. There was much information available on the Web regarding digital archiving, but it gave few specifics. Many of the documents, as with the traditional published literature, emphasized the issues and challenges of digital archiving, rather than documenting actual systems experience. However, the literature search provided several valuable bibliographies in digital archiving and electronic publishing, and helped to identify additional experts and projects.

Contacts were also made with numerous information organizations. These included national and international library organizations (International Federation of Library Associations, ASLIB, Association of Research Libraries, Council on Library and Information Resources, the Research Library Group, the Coalition for Networked Information, the Federal Library and Information Center Committee, the Online Computer Library Center, and the Corporation for National Research Initiatives), publishing and database producer organizations (National Federation of Abstracting and Information Services, Society for Scholarly Publishing, Association of American Publishers), national libraries (British Library, the U.S. national libraries of Agriculture, Medicine, and Education, and the Library of Congress), electronic records management and archive organizations (Archimuse, National Archives and Records Administration, UK Public Record Office) and digital libraries (Digital Library Federation, Los Alamos National Laboratory's "Library Without Walls", and the California Digital Library).

In addition, the Principal Investigator attended three conferences. A meeting of the U.S. Department of Agriculture Digital Publications Preservation Steering Committee was attended on February 19, 1999. This meeting provided insights into the issues from the point of view of operations staff in one area of the scientific discipline. In addition, the Principal Investigator was asked to give a presentation on the effort being undertaken by ICSTI and CENDI. The second meeting was a U.S. National Science Foundation *Workshop on Data Archival and Information Preservation* held in Washington, D.C. on March 26-27, 1999. The Principal Investigator participated in the general sessions, and in specific discussions about the requirements for digital archives and policy issues surrounding archives. She also described the ICSTI/CENDI effort when the discussion turned to the question of model projects and best practices. Finally, the Principal Investigator attended the wrap-up sessions of a NISO/CLIR/RLG *Technical Metadata Elements for Image Files Workshop*, held in Washington, D.C. on April 19, 1999.

Since this field is relatively new, and, often operators are not researchers and writers, it is not surprising that the majority of the information came from personal contacts and word of mouth. (A list of contributors and additional contacts is presented in Appendix D-1 and D-2.) From the literature searches and other expert contacts, an additional 16 projects were identified for initial review.

Based on the initial review of the 35 candidate projects, 19 projects were selected for more

detailed review. Follow-up discussions were conducted via e-mail or in person for 15 of the selected projects. The discussion outline is provided in Appendix A-2. In four cases, the organizations did not respond to e-mail and telephone requests for information, but sufficient information was found via their Web-sites and additional literature.

Highlighted Projects

The purpose of the initial survey was to identify operational and prototype projects that could be considered noteworthy, innovative, or cutting edge. It was important to "weed" through the responses to identify those projects which should be highlighted. The criteria for selection included:

- adherence to the ICSTI/CENDI definition of digital electronic archiving, i.e., that the original was published in digital form or that the digital archive will be the sole or primary archive and that the purpose of the archive is long term preservation and reuse
- innovation and "cutting edge" approaches in areas such as metadata standards, storage technologies, intellectual property rights management, cost/resource models, policy development, etc.
- degree to which the system is operable
- inclusion of scientific and technical information
- data type; since ICSTI/CENDI were interested in a variety of data types applicable to science and technology an effort was made to ensure coverage of major data and document types in the sciences

The following table highlights the 19 selected projects by key characteristics. More complete descriptions are contained in Appendix C. The majority of the information concerning possible models, best practices and costs was developed from these projects which were considered to best meet the criteria for "model" projects outlined above. However, additional projects that were not reviewed in as much detail are used throughout the report to show trends and comparisons in particular areas.

Highlighted Project	Brief Description	Special DEA Characteristics
American Astrophysical Society (US)	Learned society archiving its own journals and also linking to a larger international system of astronomy literature.	Collaborates with other astronomy societies and government organizations to maintain complete linked access to the astronomical literature, including an archive of core literature for the last 150 years. Found money for major system migration will likely be covered by current operating costs rather than special escrow fund.
American Institute of Physics (US)	A learned society in physics which archives the electronic journals and supplementary material.	Extensive licensing agreement information for customers. Provides archiving as a service to member societies who publish.
Atmospheric Radiation Measurement Program (US)	Data center that stores data and metadata generated by this DOE program which measures sunlight, meteorology, clouds, temperature, water vapor, etc..	Large volume of data ingest (6-8 GB per day) and use (1-2 GB per day).
Carbon Dioxide Information Analysis Center (CDIAC) (International)	A subject-specific World Data Center that monitors carbon dioxide levels.	Meets the criteria for a World Data Center. Requires agreement for perpetual care of datasets.
Distributed Object Computation Testbed (US)	This pilot project at the San Diego Supercomputer Center and numerous other locations is sponsored by DARPA and the US Patent and Trademark Office.	The Distributed Object Computation Testbed (DOCT) has created a testbed system for handling complex documents on geographically distributed data archives and computing platforms. The technologies should apply to the information needs the US PTO and other U.S. federal agencies. Technologies include replicated archives, redundant communication paths and advanced database technologies to access heterogeneous databases, in a secure environment.

DITT (Defense Information Technology Testbed) (US)	A project of the U.S. DoD. Effort funded by the National Technology Alliance.	Archiving multimedia objects including video from Unmanned Aerial Vehicles, still imagery, transcribed text, and audio voice-overs by pilots. Original material received from Joint Analysis Center in the UK and information is stored and preserved at Ft. Leavenworth, Kansas, US in a multimedia data warehouse system.
Electronic Publications Preservation Project/Electronic Journal Collection (Canada)	EPPP was a pilot project to preserve electronic journals published in Canada. The Electronic Journal Collection now incorporates them in the normal workflow.	Incorporated preservation of Web documents into regular flow of deposited material. Addressed intellectual property and document extent issues. Recently published guidelines.
Environmental Information Management System (EIMS) U.S. Environmental Protection Agency (US)	The EIMS system is providing an EPA-wide information management environment centered around distributed databases.	Developing a structure in which archived datasets can be automatically ingested into Oracle databases for integration and reuse with other databases, using tools as they evolve.
EVA - the acquisition and archiving of electronic network publications (Finland)	A joint project of libraries, publishers and expert organizations led by the Helsinki Univ. Library-Finnish National Library and funded by the Finnish Ministry of Education.	Harvesting and archiving digital information relevant to Finland. Connections to international and regional standards. Emphasizes added value through links and interactive e-journals. Published collection guidelines.
HighWire Press (International)	Stanford Univ. Library program to support small learned societies in electronic publishing.	Provides archive along with other electronic publishing services.
JSTOR (International)	Third party repository originally funded by the Mellon Foundation. Now a non-profit organization.	"Final resting place." Often required by library consortia as trusted third party depository. Well-developed, tiered cost structure for access. Archive is built both from current electronic submissions and scanning of paper backfile.

Kulturaw3 Royal Library, National Library of Sweden (KB) (Sweden)	Royal Library, National Library of Sweden project to test methods for collecting, preserving and accessing Swedish electronic materials, including periodicals, static documents and dynamic document with links.	The approach is to make the capturing as automatic as possible. Robots have been tested which collect publications based on the .se extension, location of the Web server in Sweden even though it has a .com extension, and foreign produced pages with a Swedish connection, e.g., translations of Swedish literature. Material would be collected and downloaded to the KB server routinely without review or selection. For electronic periodicals, a method similar to that of PANDORA is used to monitor and harvest every issue. Also testing hierarchical storage architectures, off-line archive storage methods, and metadata requirements. At least two runs of the robot have been done. Policy that requires deposit of electronic journals.
Long Term Ecological Research Network (LTER) (US)	Federated centers for ecological information.	Working on a heterogeneous network approach with the National Center for Ecological Analysis and Synthesis (NCEAS) and the San Diego Supercomputer Center.
NASA Distributed Active Archive Centers (US)	Federation of expert centers for global change information	Mature infrastructure for archiving datasets and providing customer support, information products and tools.
National Digital Archive of Datasets (UK)	Centralized data center for depositing government datasets within the UK Public Records Office. PRO has separated responsibility for datasets from "office documents".	Working on standards for documentation to be supplied with datasets.
Natural Environment Research Council (UK)	Data center for environmental science in the UK.	150+ year history for some of its data collection centers. Federated data centers with extensive data management guidelines.

OCLC Electronic Journals Project (International)	Online bibliographic utility, which provides a variety of services to libraries including copy cataloging, serials cataloging, search utilities, etc.	Provides e-journal archiving as a service to member libraries and makes them available via the OCLC network. Also providing links between local library journal holdings and the full text, so that the article level bibliographic data is retrieved through an OPAC search.
OhioLINK Electronic Journal Center (US)	Library consortia of various library types in the state of Ohio.	Agreements with publishers require electronic resources to be archived by the EJC.
Preserving and Accessing Networked Documentary Resources in Australia (PANDORA) (Australia)	National Library of Australia project to archive Internet-based Australiana, including Web sites, electronic journals, etc.	System for automatic harvesting of Internet with a "gathering schedule." Guidelines for selection of all types of electronic objects, including ephemera.

The highlighted projects cover six countries (U.S. (9), UK (2), Canada (1), Australia (1), Sweden (1) and Finland (1)). Four organizations are considered to be international in scope, because their funding sources and scope are not bound to a particular country. The projects come from a number of sectors including government scientific and technical programs, national archives, national libraries, learned society and commercial publishers and other research institutes. The major object types in the sciences are also included – electronic journals, technical reports, numeric data and patents. The major data types are also included -- text, data, images, video and multimedia.

General State of the Art/Practice

The issue of archiving digital objects brings together several normally diverse communities -- archivists, records managers, librarians, data center managers, and data producers. There is so much activity among various groups that it is difficult to encapsulate the general state of DEA. There are numerous groups working on the issues, for example library organizations (Council on Library and Information Resources and the International Federation of Library Associations), archivists (Society of American Archivists and Archimuse), and numeric data collectors (Consultative Committee for Space Data Systems). Funding has been provided by the European Commission (the Committee on Telematics and Telecommunications, through the DLM), governments and government agencies (Defense Technical Information Center, National Library of Medicine, U.S. Department of Agriculture), public archives (the U.S. National Archives and Records Administration and the UK Public Records Office), depository libraries (the National Library of Australia, the National Library of Canada, the U.S. Library of Congress, the British Library), private grants (Mellon Foundation, Long Now Foundation), the U.S. National Science Foundation (the U.S. Digital Library-1 and 2 research initiatives and a recent National Science Foundation Workshop on Digital Preservation), and individual organizations (Getty Information

Institute, the Arts and Humanities Data Service, and the Research Library Group).

Within the traditional archives and records management organizations there has been much interest. While many of these activities focus on administrative and government documents and datasets, they also include documents and datasets resulting from government funding of scientific and technical research. The Society of American Archivists has issued a statement on the preservation of digitized reproductions (www.archivists.org/governance/resolutions/digitize.html). A joint statement concerning electronic records management has been issued by the public records archives of the US, Australia, the UK and Canada (www.pro.gov.uk/recordsmanagement/eros/ercollab.htm). The UK, Australia and the European Union have significant electronic records initiatives underway. It is significant that the governments with extremely active electronic records initiatives, such as Australia, the UK, and various Nordic countries, also have some of the most advanced activities in more traditional library areas.

There are several major groups that have done significant background research in the area of DEA in relation to digital libraries. These include the Research Library Group, the Arts and Humanities Data Service, and the UK Online Library Network (particularly the e-Library Programme). While the RLG has members who are scientific and technical librarians, the majority are in the arts and literature. The AHDS is responsible for much of the effort in the arts and humanities for higher education in the UK. While the focus of this project is not on the arts and humanities, many of the reports and white papers sponsored by these groups can inform the discussion of digital archiving for other disciplines. Therefore, these documents are used heavily in this review and the results and ideas are extrapolated to the scientific and technical arena.

The state of DEA is interesting in that the "cutting edge" projects may not be in the physical or social sciences, but in the humanities. There are numerous projects that have as their basis literature, art, and cultural heritage. The latter is a particular motivator for governments who feel that they are losing a generation of culture, because it is published on the Web and then disappears. There are several major projects in this area including SCRAN (Scottish Cultural Resources Access Network) and the AMICO Project (Research Library Group, and the San Diego Supercomputer Center). Some of these projects address the most complex type of electronic archiving -- that of multimedia objects.

Identified Organizational Models

.

The highlighted projects were analyzed for commonalities that would identify operational models for DEA. The approach taken is an organizational one, emphasizing the role of the archive center in the information chain. The analysis is loosely based on the previous work sponsored by the Arts and Humanities Data Service (AHDS) (Beagrie and Greenstein, 1999.), which provides a framework for digital archiving based on the information life cycle -- creation, management, dissemination and storage. Rather than the six organizational structures (data banks, digitizers, institutional archives, academic archives, legal deposit libraries, and funding agencies) identified

by AHDS, we have identified five major models (data centers, institutional archives, third party repositories, publication service providers and legal depositories). These models are based on differences in the information flow, responsibility and ownership of the data, and the economic model. In each case, the distinct characteristics of the organizational model are described, along with important information concerning intellectual property and economic issues.

Data Centers

The data center model is the most mature within the scientific community. Some of these data centers have existed in one form or another since the 1960's. The role of the data center is to store and locate upon request the data that the creator or producer provides to them. Generally, the operational archives of this type in the sciences collect numeric data, with text limited to documentation files. The data deposited with the center may be created by the center itself or deposited by others who are partners in the particular mission, but the acquisition and collection policies are determined by the sponsor. In addition many data centers provide services to a particular user community as a means of disseminating the information more broadly, and also, in some cases, raising revenue to support the activity. Use of the information is often key to the centers' missions, and so the data centers are often involved in the development of summary products from a single dataset or the integration of multiple datasets, the creation and distribution of software for use with archived data, and customer service to their particular user groups.

Many of these centers support large-scale, global data collection programs in the Earth and environmental sciences, including climatology, meteorology, and global change. Significant data centers also exist in molecular biology, genetics and biochemistry. These data centers often categorize themselves as *active archives*, meaning that the data is continually reused and added to. The benefit of the archive is in its continuous reuse, modification, and integration. The size and longevity of the archive are its two biggest assets.

Three subcategories of data centers have been identified in this study -- centralized, federated and cooperative. These categories are based on the degree to which the archival storage and management responsibility is distributed across one or more sites.

The Centralized Data Center

The centralized data center acquires digital objects from other sources within its discipline or region, taking over sole responsibility for their preservation. The acquisitions may be based on a network of affiliated organizations, but the archive itself is centralized.

The UK National Digital Archive of Datasets (NDAD) is an example of a centralized approach. Under contract to the Public Records Office (PRO), the University of London Computer Center and the Library of the University of London have developed an accessible archive of government datasets. The PRO continues its responsibility for selecting what should be held in a long term archive, then it is the responsibility of the NDAD to get the dataset into its system, to catalog it to

standard, to transform it as necessary, and to preserve it. In some cases, the NDAD also provides the software needed to use the dataset. While many of the datasets are administrative in nature, there are several that deal with environmental monitoring or hydrology data.

NDAD has a single infrastructure which supports the archiving services of various forms (Ashley, Personal communication. 1999). In a recent article in the NDAD Newsletter, Kevin Ashley, Project Manager, emphasizes the importance of an integrated archive. He points to the fact that by having scientific datasets that span multiple locations, the whole (which is still growing) is greater than the sum of the individual parts. This is also true of the CDIAC, where datasets have been integrated that uniquely cover natural events for over 20 years.

Another example of a centralized archive approach is that of the World Data Centers (of which CDIAC is one). To become approved as a WDG an agreement is made to acquire and collect data on a technical area, make it available for active research, and than maintain it in an archive. If a center should go out of business, part of its responsibilities are to ensure the data is transferred to another institution for preservation.

The benefits of a Centralized Data Center:

- increased control on the part of one archiving organization
- easier adherence to archival standards
- easier integration of data from various datasets

The challenges of a Centralized Data Center:

- funding is generally tied to a major sponsor who drives policy and visibility of the preservation of material. To the extent the mission is active research there is a question of what will happen to the older data.
- may leave some key information out of the loop for users because the producer or owner cannot meet the requirements for central deposition
- possibly more difficult for the central organization to react to changes in hardware and software technologies and the needs of the user community
- requires transfer of data or redundant storage by two organizations, which may result in issues of validity and ownership

Federated Data Centers

The federated data center model consists of a series of distributed organizations that take responsibility for a particular area of expertise. That area may be built on subject, geography, or organizational mission. Each node in the network has responsibility for a defined portion of the science, but overall policies for preservation and access are established at a central management level. They "collectively provide a physically distributed but logically integrated database.

The most prominent example of this model is the federation of NASA's EOSDIS Distributed Active Archive Centers (DAACs). The eight DAACs support the Earth Observing System (EOS), which has responsibility for the long-term global change research program designed to improve understanding of the Earth's interrelated processes involving the atmosphere, oceans, land surfaces, and polar regions. These data centers are hosted by geographically dispersed institutions, including government installations, such as NASA's Langley Research Center and NASA's Goddard Space Flight Center, academic institutions such as the University of Colorado at Boulder, and by contractor operated sites such as the Oak Ridge National Laboratory.

Each center processes, archives, and distributes EOS (Earth Observing System) and other NASA Earth Science data in a narrow area of the discipline (land processes, upper atmosphere, snow and ice, biogenochemical dynamics, hydrologic cycle, etc.). Each data center provides services tailored to the specific needs of its individual discipline and user communities. Together they provide over 900 data sets and coordinated services (access, redaction and summarization, analytical tools, customer service) to support interdisciplinary Earth science research. (ivanova.gsfc.nasa.gov/daac/)

Access to the entire system is provided through the Earth Science Search and Order System hosted at the Goddard Space Flight Center, using Goddard's IMS Web Gateway. Users are able to search for and order from any of the DAACs through a single search. The IMS provides both Web and graphical user interfaces to accommodate a variety of user computing environments ranging from desktop PCs to sophisticated graphical workstations.

The Natural Environment Research Council (NERC) data centers in the UK are organized in a similar fashion. The seven centers are housed at universities and research institutions, based on their expertise and infrastructure capabilities. The infrastructure developed by NERC supports not only the archiving of the digital data, but its active access. This must be in place in order for an organization to be considered a data center in the NERC federation. The integration of the distributed databases is achieved through common data formats and a well-defined data management policy that is shared across the federation (www.nerc.ac.uk/environmental-data/). In addition to the active life cycle management for new datasets, NERC continues to incorporate legacy datasets into its collection. George Darwall, head of the NERC regrets that the regimen of life cycle management was not in place for the 150+ years for which some of the organizations involved have been collecting data. (Darwall, Personal communication. 1999)

The Long Term Ecological Research (LTER) Network fosters the synergy of information systems and scientific research toward the goal of promoting ecological science (www.lternet.edu/documents/Reports/Data-management-committee/1995-DM-committee-report/im_1995_report.htm). The LTER sites collect and archive ecological data. Some sites also archive related textual material such as proposals, theses, papers and research summaries. The LTER Data Catalog contains over 2,000 entries (www.lternet.edu/DTOC). The LTER sites are developing a Networked Information System, a "distributed, LTER-wide information system using a modular approach, while maintaining and building on present

functionality." (Porter, Personal communication, 1999.) The sites have a relatively high degree of autonomy. For example, the migration for hardware and software is the responsibility of each site and at any given time, one or more sites are undergoing significant upgrades. However, the data managers meet regularly and share best practices and common concerns.

The economic model for federated data centers generally includes a combination of earmark funding via contract or grant from a sponsoring agency or organization, and fee for service. Both the NASA DAACs and NERC charge for the datasets depending on the use to which they will be put. Much of the DAAC data is government produced, and, therefore, the charge is solely for the formatting and distribution, not for the original collection of the information. In the UK, NERC aims to provide "inexpensive access" to those researcher who will advance the knowledge of the field but not for commercial gain, and who will publish their results in the open literature. Data may be supplied to these users either free of charge, at a nominal handling fee, or at a discounted rate. Revenue from the commercial use of NERC's data is used to offset the cost of the collection and long term data management of the archive.

The question arises how well this model holds for other data types? While these centers archive terabytes of numeric data whether from ground or remote sensing instruments, the data is fairly homogeneous and simple binary or ASCII data streams.

The benefits of the Federated model include:

- integrated databases that can work together to support the mission
- backup for the provision of customer services that span the centers
- provision of special tools and services by discipline or user community
- power to create policies

The challenges of such a model include:

- funding may still be tied to a major integrating sponsor who drives policy and visibility of the preservation of material
- the integration of distributed databases is difficult to maintain, particularly as new software and data management approaches appear, since it requires consensus across multiple archives
- it may be difficult to add new centers of expertise as the communities and disciplines change
- it may be difficult to develop data structures, interfaces, and information management policies that are scalable to all scientific and technical disciplines and object types (both NERC and DAACs focus on relatively small areas of the total of science; and on a single format type)
- it may be difficult to use this model for archives where the designated user communities are numerous and very fragmented, because one of the main goals is to provide standardization for a specific user community

- ensuring funding for the coordination effort that must still go on to ensure compatibility across the federation

Cooperative Data Centers

The challenge of maintaining integration across federated data centers has led to the idea of developing a looser federation of centers based not on homogeneity, but on heterogeneity. In a recent technical report on the issues surrounding the archiving of Earth science data, Bruce Barkstrom of the Atmospheric Sciences Division at NASA Langley Research Center, argues the federated DAAC model cannot be maintained into the future and that the ability to search across heterogeneous databases, including legacy databases, is critical to access in the future (Barkstrom, 1998). Mr. Barkstrom promotes the idea of cooperative data centers. This approach espouses that the current data centers will evolve into heterogeneous data centers that exchange data. Barkstrom, who acknowledges that this level of cooperation will take a long time to implement, states that "this vision suggests a future that contains individual data centers that cooperate to provide services that are more helpful than any could provide alone. This vision does not require a single homogeneous approach." Issues such as the long-term archival requirements, the data structures for archival holdings, and the user views of the data will differ by discipline and by user. This calls for extreme flexibility, while requiring extensive documentation and adherence to standards in such documentation. It is noteworthy that the standards for documentation submitted along with datasets is a current project of the UK National Digital Archive of Datasets.

The cooperative center approach is at the heart of a budding consortium of ecological data producers. This approach "recognizes the highly distributed nature of ecological data as well as its extreme heterogeneity in structure and content." (Jones, Personal communication, 1999) This project will federate information sources through a distributed data network including the Long Term Ecological Research (LTER) Network, various field stations and laboratories, the National Center for Ecological Analysis and Synthesis (NCEAS) at the University of California at San Diego and the San Diego Supercomputer Center (SDSC). This national network is in the design and early prototype stage, but it will involve the use of highly structured metadata in XML to facilitate integration, access, and exchange of ecological data. Most of the work to date has been done at the individual participating institutions -- LTER (the information system), NCEAS (structured metadata), and SDSC (a distributed heterogeneous data system called the Storage Request Broker). Several proposals have been submitted to increase the funding and the pace of this project.

The proposed benefits of a cooperative data center model include:

- increased flexibility and autonomy among the participants
- easier incorporation of new centers because they do not need to meet such stringent guidelines for incorporation
- may more adequately address the heterogeneity of content and data types in certain

disciplines such as ecology
- may allow broader areas within and across disciplines to be networked, based on a core standard with extensibility for what is unique about each discipline
- since multiple organizations are assumed to be involved, there may be a broader and more stable funding base – if one organization drops out the others carry on

The challenges of the proposed cooperative data center approach include:

- reaching the minimal standard for interoperability – achieving the balance
- developing standards for the documentation that is required to adequately describe the datasets and any required software
- integrating the tools and content while not confusing the user community
- ensuring funding for the coordination effort that must still go on to ensure interoperability

The economics of the data center model is fairly simple. The owner or producer of the information funds the data center or network of data centers to store and make the data accessible to the user community. In most cases, there is some community (whether the employees of a particular organization, members of a society, or the general public) who have access to the archive in an online environment for no cost. Other user groups may be charged for access in order to recover some of the costs of the archive. Charges may also be levied for special services, such as customized datasets and formats, or for commercial uses of the data. Unfortunately, data centers are generally supported by specific programs or projects. While this provides a focus for the data collection, user community interaction and creation of added-value information products, it also makes the center dependent on short-term rather than long-term funding.

The Institutional Archive

Institutional archives are departments or branches of an institution that collect and preserve the intellectual capital for that institution. These institutions include publishers, data producers, societies, cultural organizations, government agencies, academic institutions and industries of various types. The role of the publisher as archivist is covered in more detail in the section on Life Cycle Managers and Their Roles -- Publishers.

While in many institutions, this type of corporate archive is more interested in preserving the history of the institution, including changes in ownership, directorship, business practices, etc., many of these organizations are also involved in scientific and technical information. The expressions of this research may differ depending on the type of organization.

The “cutting edge” research in science and technology from academic institutions may take the form of datasets, software or other objects. The results may be expressed in conference papers, laboratory notes, contract/grant reports, preprints, or formal monograph or journal article

presentations. Faculty members are also beginning to develop their own Web sites, which may include the results of their research, in additional to biographical information. The results of student's work may be presented in these forms as well, but also in theses and dissertations.

Particularly prevalent among universities are examples of digital dissertations and theses. In addition to the individual universities, there is a federation of universities --- the Networked Digital Library of Theses and Dissertations (NDLTD). The individual theses and dissertation servers from other universities that are linked in this federation continue to grow, with many universities in Europe joining. Driven by the advances in distributed information processing, we see the push here, as with federated data centers, toward a loose federation based on commonly accepted standards. Joining requires the installation of certain server software and text submission in PDF. Much time has been spent on the social and organizational aspects of theses and dissertation deposit, but there is no formal plan for digital archiving, other than routine backup and recovery procedures. (Fox, Personal communication. 1999) It is not a requirement for NDLTD partnership that there be a plan for long term preservation. This has served as a model for similar projects at the Royal Institute of Technology in Sweden and some Australian universities are investigating a distributed archive of research theses, modeled on this approach.

There are many examples of institutional archives within industry, and although there are very few that focus on digital information, this is the most rapidly growing area, particularly with the impetus from "knowledge management" trends. In a recent NSF Workshop on Digital Preservation it was noted that the major impetus for formal electronic archives within an organization (or moving from paper to electronic) is mandate or regulation (Busch, Personnel communication, 1999). This is particularly true within the pharmaceutical industry where digital archiving is a required follow-on to electronic submission to the U.S. Food and Drug Administration. Smith Kline, for example, has a significant program in this area, which must, in the short term, cope with a variety of word processing, database and modeling formats. (Brunone and Roberts, Personal communication, 1999). Other scientific industries with similar needs include the chemical and petroleum industries.

It appears that industry archives are likely to grow in the future, particularly if the benefits of knowledge management systems develop as envisioned. In connection with intranet development, many companies are purchasing software to support knowledge management and consider the information which is now being saved to be part of a long-term archive that will support decision making in the future. While the current emphasis may be on the document management and integration of disparate information sources across the enterprise, as the corpus of material builds, it is likely that more attention will be paid to digital electronic archiving issues in this sector. However, juxtaposed with the benefits of digital preservation by industry is the concern among the corporate legal community that preserving this information may result in unwelcome legal actions and outcomes in the future.

There was insufficient time in this study to pursue extensive examples of DEA within the proprietary environments of scientific-related industries. However, it is likely that examples exist

that would inform the discussions about the preservation of particular data types and uses or archival information by various user groups.

The benefits of an Institutional Archive include:

- provides a repository of cutting edge research
- can be based on a formal organization structure where rewards and incentives can be applied to digital archiving
- the archive is organizationally closer to the originator of the information which may make communication on format, migration and reuse easier
- the archive is organizationally closer to the funding source of the information which may give the archive more lobbying ability

The challenges of an Institutional Archive include:

- in many institutions the incentives are not in place, based on the culture that recognizes only formal printed publications or near-term information exploitation
- depending on the type of institution, the digital archive as an overhead item may be considered expendable when budget cuts are necessary
- depending on the size and primary business of the institution, it may be ill equipped to handle any activities beyond the simple storage of the initial DEA submission, leaving the data preserved but inaccessible

Third Party Repositories

An newly emerging archival model is the Third Party Repository. This is an organization other than the originator or institution that owns the object (publisher or other institution archive) that archives and preserves objects from one or more originators or owners. Two types of third party repositories have been identified: repository management agents and publication service providers.

Repository Management Agents

Repository Management Agents serve as trusted third party repositories, but do not serve any other function in the value-added chain. They are acting as agents for the learned societies, the publishers, or the creators. This new "organization" provides a safeguard for the other points in the system by providing access to the digital work should the publisher or producer of the information determine that they can no longer archive the material or if they go out of business.

This is the type of service provided by JSTOR. JSTOR focuses on journal literature, both current electronic versions and paper backfile conversion. The project was originally funded by the Mellon Foundation, but is now incorporated as a not-for profit organization. Phase I of the JSTOR Project scanned and archived 117 journals in 15 humanities and social science disciplines.

Phase II, which was recently announced, is a general science cluster beginning with agreements with the American Association for the Advancement of Science (AAAS) and the Proceedings of the National Academy of Sciences to archive their electronic journals and convert back issues from paper. This will include over 100 years of scientific literature from AAAS and PNAS issues dating from 1915. (It is interesting to note that AAAS, as a publisher, is also working with multiple repositories. In addition to JSTOR, AAAS is distributing its electronic version, *Science Online*, on its own Web site and through the HighWire system.) Elizabeth Bennett, head of the JSTOR's Princeton Production Facility acknowledges that archiving scientific literature is different from the humanities in terms of the complexity of layout and the amount of color and graphics.

OCLC's Electronic Journals Project also provides third-party archiving. OCLC takes a publishers data and makes it available to member libraries that have a subscription to the electronic journal, either directly with the publisher or through a jobber. OCLC currently has over 2200 journals from 46 publishers, of which 1500 are mounted. What differentiates the OCLC service from similar services provided by jobbers is that OCLC is committed to archiving the journals forever. The publisher must agree to send a copy of the appropriate electronic issues in PDF to OCLC if it can no longer provide access or if it goes out of business. In most cases, the publishers provide the electronic copies immediately, because OCLC as an online utility is better equipped to handle the network resource issues, such as multiple simultaneous users. (Hearty, Personal communication. 1999.) However, there are still some publishers that have their own online systems and for which a pointer is provided in the bibliographic data that links to the publisher's online system. OCLC retains a database indicating the year to which the library has subscribed to each particular journal, so that access is given only to those issues for which the library has subscribed. Users pay a small access fee. As users come increasingly to OCLC for "one stop access", OCLC is reviewing its business models and fee structure for this service. The recommendations will be shared with the OCLC User Council and Board later this year.

Another notable third-party agent is The Internet Archive (www.archive.org) created by Internet guru, Brewster Kahle (www.sciam.com/0397issue/0397kahle.html). The Internet Archive takes snapshots of the Web and preserves all Web pages, newsgroups, ftp sites, gophers, etc. that are publically accessible or that have not been tagged as "off limits" through a registration form that allows owners to indicate that they do not want to be included in the archive. The collection is over 12 terabytes and contains 5 separate snapshots of the Web. The Internet Archive is now in the process of gathering resources through its commercial entity called Alexa Corp., to make the archive accessible to the public via a simple Web interface. They expect to have the interface developed within a year or so. (Mack, Personal communication. 1999.) While Dr. Kahle's approach has received much press and discussion within the community, this approach requires significant resources and it is unclear what to do with the snapshot once you have it. However, a copy of one of the snapshots has been ordered by the US Library of Congress for research purposes.

The examples of third-party repositories that have been implemented to-date have a user-oriented

economic model. The third-party receives the archive from the producer, but in the case of OCLC and JSTOR, the expense is paid by the user or library. In the case of JSTOR, there is a multi-tiered pricing schedule based on the size of the institution. The pricing is also divided into an annual fee for the maintenance of the archive and a one time start-up fee. This provides specified users at the subscribing institutions access to any item in the archive for the subscription period. It should also be noted that neither of these operational systems are commercial in nature. Both OCLC and JSTOR are incorporated as not-for-profit organizations. There is no example yet of a truly commercial digital electronic third-party archive. This may indicate that the industry is not mature enough or that commercial entities do not envision sufficient profits from a business that provides archived digital materials.

The biological sequence data bank is an unusual type of third party repository archive, because it is a public/private collaboration for which no fees are charged. It reflects the efforts of an entire scientific community. As the sequences to be printed in paper copy became more complex, there was a move on several fronts, particularly among the biomedical publishers, to have the sequences deposited and to simply print an identification number that would provide access to the sequence information. Many organizations including the national libraries, learned societies and biological publishers supported the requirement that sequences must be deposited with a data bank prior to acceptance of a journal article for publication. Through this institutional requirement the data banks related to protein, nucleotide and gene sequences have expanded and become valuable resources particularly for computer manipulation. While some of these data banks have intellectual property issues associated with them (the Online Mendelian Inheritance in Man from John's Hopkins Univ.), some of the largest (GenBank and the Protein Sequence Data Bank) are in the public domain. Fees are not charged for depositing, searching or downloading the information. There are some members of private industry that take periodic copies of the GenBank data to load in-house, in order to ensure privacy relating to the kind of searching that is being done on the data bank. (Benson, Personal communication. 1999.)

The benefits of a Repository Management Agent are:

- sharing of resources and costs for smaller publishers/producers
- may be necessary in order for a publisher to meet a library's requirement for electronic archiving of a licensed electronic resource
- organization is focused on archiving issues
- if repository is focused on a particular data type it can provide specific tools for that type

The challenges of a Repository Management Agent are:

- ensuring that the third party can be trusted
- tailoring the storage and access when the repository is working for multiple publishers/producers

Publication Service Providers

These organizations provide publishers with a variety of services including design, development, distribution and marketing of electronic journals. These publishing services may also be jobbers, brokers, agents, or network providers. They are involved in these traditional aspects of the scientific information chain, and have enhanced these services by providing avenues for electronic journal archives to be created. The benefit from this approach is that these organizations (EBSCO, Blackwell, Dawson, Ovid, Swets) act as "E-Journal Consolidators" (Okerson, 1999.) providing access to multiple journals as a single collection. However, few of them have acknowledged that they will take on the responsibility of long-term archives.

An exception is HighWire Press (intl.highwire.org/) , one of the largest of these publication service providers turned archives with 110 journals online as of March 1999. HighWire is similar to JSTOR, but it has publishing responsibilities for many of the journals that it archives, and has focused on journals in science, technology and medicine. The majority of its partners are scientific societies. Stanford University founded HighWire as a department within Stanford in 1995 over concern that these societies would not be able to transition to the technologies needed for more advanced scientific communication in a networked world. HighWire has the role of "partner, agent of change, and advisor."

Under the guidance of the development teams which include scientists, librarians and publishers, HighWire's approach to online publishing of scholarly journals is not just to mount electronic images of printed pages, but rather to add links among authors, articles and citations. HighWire has also developed advanced search capabilities, provided high-resolution images and multimedia as appropriate, and works toward a more interactive electronic journals environment.

Unlike, JSTOR, HighWire does not appear to have as much interest in providing a complete backfile archive of a particular journal title. It does not do extensive scanning of backfile materials. The span of issues available via HighWire depends on the particular journal. Several of them have electronic versions dating back to the early 1970's. The two services appear to be complementary in that J-STOR focuses on the older material and HighWire focuses on providing innovative services into the future. For example, *Science* on HighWire dates to 1995, while *Science* on JSTOR is planned to include all 100+ years when the paper issues are scanned.

Another example of a publication service provider that also provides archiving services is ADONIS (www.adonis.nl/) . This organization, which started as a collaboration among several publishers to test the provision of electronic journals on CD-ROM, is now owned by Blackwell. They provide an archive for over 60 scientific, technical and medical publishers.

Providers of preprint server systems may also act as long term archives. Preprint servers are systems, including storage, access and presentation interfaces, that provide access to pre-publication materials. In the case of certain preprint servers, these have expanded to include

material past the prepublication stage of the life cycle. These systems support the review (peer or informal comment process), bibliographic access and subsequent archiving for preservation. Preprint servers may be organized around the discipline or the institution.

The most famous preprint server is Paul Ginsbarg's Preprint Server at the Los Alamos National Laboratory. This project began with the narrow discipline of High Energy Physics and then expanded into other areas including math and computer science. This preprint server is also becoming a vendor or archive repository service by serving as a host for other organizations' archives. The Association for Computing Machinery (ACM) has announced not only that its scholarly journals will be available through the preprint server, but that the LANL system will support the archiving of the history of computer science. (Arms, Personal communication, 1999).

The relationships between Publication Service Providers and other archiving entities can be very complex. For example, ingenta, Ltd., an electronic publishing service in the UK that provides online journals and databases to a consortium of academic libraries in the UK, creates the electronic journal files for small or medium sized publishers. In an agreement between ingenta and OCLC, these files are then provided to OCLC for archiving purposes (www.ingenta.com/Tfedocs/press/oclc.html). The arrangement for archiving is made between the publishing services agent (ingenta) and the third party repository (OCLC), rather than directly between the publisher and the third party.

The economics in the Publication Service Provider model is also the most complex. The business relationships are heavily dependent on the size of the community being served, the commercial value of the current and archive information, and other business relationships that may exist between the entities. For example, in a particular instance the publication service provider may be gaining sufficient revenues from the publisher or producer of the digital work that the archiving of the information is included without cost. In other cases, there may be no supportive revenue stream and the publisher must pay for the archiving service. In some cases, there will be sufficient revenues to be gained from users of the archive on an ongoing basis that the revenue is either solely provided to the archive or shared with the publisher.

The benefits of Publication Services as archives are:

- publication services understand the particular producer/creator market they support
- may be less expensive as revenues from other publication services may be used to subsidize the cost of archiving

The challenges of Publication Services as archives are:

- ensuring proper focus on the Publication Service on the long term preservation and archiving issues, when their main business may be in other services
- ensuring the longevity of the publication service
- managing the variety of intellectual property issues involved in a more complex business

model

Legal Depositories

There are generally two types of legal depositories, national archives and national libraries. These institutions generally differ in the type of material that is collected and the purpose for that collection. The responsibility of the national depository (or archive) is to document the business of government, which includes administrative documents. The libraries are generally charged with maintaining the culture, history and intellectual output of the country by collecting what is published within that country either in general or in a specific discipline (agriculture, medicine, etc.). While these types of legal depositories may be housed in the same organization, there are differences that are of interest to the archive question.

National Libraries

In addition to the publishers, there is a significant interest in electronic preservation on the part of national libraries. The national libraries are mandated to acquire, catalog, preserve and provide continuing access to the published material from their country or in support of a particular national interest such as agriculture or medicine. Many national libraries are beginning to extend their mandate into digital works.

One of the most significant projects is underway in Australia. PANDORA (Preserving and Accessing Networked Documentary Resources in Australia) is a project of the National Library of Australia. As the initial grant application states: "...the overall goal of this project is to develop and implement procedures for the capture, archiving and provision of long term access to online electronic Australian publications selected for national preservation. It will cover the full range of materials published online in Australia - including serials, newspapers and books, scholarly papers and theses, as well as unique online formats like 'homepages'. The PANDORA project hopes to provide access to future generations to an archive that represents the state of Australian online publishing from its earliest day - the incunabula period of online information - up to its most current manifestation."

Wendy Smith, who headed the project in its early phases, notes that "The hardest thing, at the moment, is ensuring that the version of any publication captured into the archive faithfully represents the online edition (www.nla.gov.au/nla/staffpaper/wsmith3.html). This has involved a careful analysis of each publication selected for archiving in order to understand both its publishing schedule and the way the information is arranged on the site. All publications currently being archived have been assessed and a 'gathering schedule' determined. This determines how frequently the online publication will be captured and a copy transferred to the archive.

Monitoring and selection of publications for the archive began in early 1996. During the first year, three publications of the initial twenty selected for the archive disappeared from the Web. An emergency rescue operation was undertaken for one of these publications, when a notice was

posted that the site would close in a few days. The other two vanished without a forward link and without informing the Library, even though the publisher had agreed to the archiving of the site. This represents a loss of around 15% of the initial material. This percent may be higher than usual due to instability in early Australian Web publishing, and it is expected that overall the loss will be less. However, the statistics point to the fact that a fully operational PANDORA archive will remedy this situation through timely capture of all relevant publications.

As of February 1999, the PANDORA archive has collected over 1,000 Australian electronic journals, magazines, webzines, e-mail fanzines, etc. They are accessible from the National Library's Web Server.

The Library is currently considering a national model for the preservation of online information. The volume of material produced in Australia is such that it is unlikely that any single institution will ever be able to preserve everything. Universities, state libraries and national research organizations may take responsibility for their own sites. The Library is also talking to state libraries about taking responsibility for state-based information such as state and local council publications.

Currently, the Library is also focusing on the PADI (Preserving Access to Digital Information) Project (www.nla.gov.au/dnc/tf2001/padi/padi.html). "PADI aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access. It focuses on providing tools, education and collaborative projects that encourage the preservation of digital information." PADI has developed a Web site with significant links to other preservation projects, particularly among national libraries. The NLA believes that the outreach and training provided through the PADI activities will reduce the complexity and perhaps the resources needed to continue the PANDORA archive.

A prototype project similar to PANDORA was conducted at the National Library of Canada. The Electronic Publications Pilot Project (EPPP) was conducted from June 1994 to July 1995. Its aim was to "identify and understand all the challenges associated with acquiring, cataloguing, preserving and providing access to Canadian electronic publications" (collection.nlc-bnc.ca/e-coll-e/ereport.htm). For the pilot project, the EPPP team used a small number of Canadian electronic journals and other representative publications freely available on the Internet. Based on the report, the results of this and other electronic publication pilot projects have been mainstreamed into the regular operations of the NLC. The NLC Electronic Collection incorporates formally published Canadian online books and journals (collection.nlc-bnc.ca/e-coll-e/index-e.htm). These publications are being acquired, catalogued, and permanently stored at the NLC. Public access is provided on the Internet through the Web. Catalogue records for Electronic Collection titles, including the Uniform Resource Locators (URLs), are available from the NLC's online public access catalog.

National Archives

The U.S. National Archives and Records Administration (NARA) has an Electronic Records Center which is responsible for the archiving, preservation and access to U.S. government records that exist only in electronic form. Generally, this center is not concerned with electronic copies of information that have already been deposited in paper. According to the center's Web site, "The records held by the Center for Electronic Records are witnesses themselves to the evolution of computer technology - our earliest records were created as early as World War II and reflect punchcard technology in use since the 1880s. However, most of the electronic records at NARA date from the 1960s, and number over 30,000 files. The scope of the holdings is quite diverse - as diverse as the activities and interests of the Federal Government itself." (www.nara.gov/nara/electronic/) NARA provides guidance to U.S. federal records managers through its publication "Managing Electronic Records, National Archives and Records Administration Instructional Guide Series" (gopher://gopher.nara.gov/00/managers/federal/publicat/elecrecs).

NARA has also funded a project at the San Diego Supercomputer Center (www.npaci.edu/DICE /nara/) to develop the architecture for a persistent electronic archive. Using Usenet messages, a corpus of word processing documents, and a series of data collections, SDSC is prototyping an architecture that will handle heterogeneous data types in collections. It uses metadata and a container architecture to store digital objects separately and recreate the collection based on the metadata. Aspects of this project include ingestion of large data collections, the metadata needed to recreate collections, the managing of heterogeneous data collections, the architecture for laying out the collections on storage media, interface design issues, and performance and cost measures. The prototype system is managing approximately 1 million records, but the project calls for an architecture that is scalable to 40 million records.

While NARA has responsibility for federal records of all kinds, the Federal Depository Library Program (FDLP) is more limited in scope. This program ensures that U.S. federally funded research publications are made available to the public through deposit of materials by agencies at approximately 1,400 federal depository libraries, most of which are housed at large academic research libraries. The mandate for the FDLP process is through Title 44 of the U.S. Federal Code, and is supported by the government printing procedures that require agencies to procure certain print services from the Government Printing Office and notification to GPO of electronic publications.. The FDLP has recently published a plan for transitioning to electronic information which includes scientific and technical information. The FDLP plan focuses on a distributed archive that relies on flexibility and a network of partnerships including government agencies and the FDLP member libraries. Generally, the archive versions of the government information will be held either at the originating agency or at the Government Printing Office, with network access available from the FDLP sites.

Under the FDLP Electronic Transition Plan, approximately 20 libraries have begun a prototype with the National Technical Information Service. (Finch, Personal communication. 1999.) In this prototype, NTIS will make available via the Internet the full text of government technical reports at no cost. For the pilot, the bibliographic database must be searched at the main library.

However, given the order number, the user can print or download the full text anywhere on campus. This prototype will provide information to NTIS about cost recovery and to the FDLP about the use and administration of these materials in a distributed environment.

The economics of legal depositories vary from country to country. However, in general, the major funding for digital electronic archiving comes from the budget appropriations for the national depository or library. Whether or not additional revenues or offsets can be obtained from customers depends on the information policy rules of the particular country. The depositories are exempted from the payment of licenses or fees for the depositing of the data.

It should also be noted that operational DEA systems at legal depositories are generally expected to be paid for from the existing budgets of the depositories. In the U.S., legislation passed in 1998 requires a substantial move toward electronic government with a subsequent impact on the NARA, with no additional funds earmarked for the handling of such records at the agency levels. In the UK, the Public Records Office developed a major electronic depository project because of the cost of increased warehouse space for paper archives. (Tombs, 1999.) The U.S. National Agricultural Library has undertaken a major development project on archiving electronic U.S. Department of Agriculture publications (Uhlir, 1998). However, the current conceptual development of the infrastructure and systems to support this must be handled within the current budget (Andre, Personal communication. 1999.) The National Library of Australia also noted that the work on archiving digital Australiana from the Internet is being done within current budgets (Phillips, Personal communication. 1999). As will be noted under the discussion of economics, lack of funding allocations lead to a basic limitation in dealing effectively with DEA challenges. It is a credit to the foresight of leaders who are making the investment based on professional commitment. The economic solutions will need to be found.

In the current situation, there are few mandates for deposition of electronic materials in depositories (particularly the depositing of electronic publications). Therefore, libraries in particular are incurring the cost of acquisition by contacting the producers or harvesting the Internet for possible archivable materials. It is unlikely that the libraries can maintain this effort, if mandates are not put in place to eliminate their information gathering costs.

The benefits of a Legal Depository are:

- mandates and regulations can require that material be made available
- may be able to require more standardization
- funding may be more readily provided for the depository than for the original program that created the electronic information
- legal and institutional infrastructures are already in place

The challenges of a Legal Depository are:

- many depositories have long histories of paper-based archives, which may create problems

when trying to upgrade guidelines and systems to support electronic records

- national depositories, unless specifically designed for a single data type or subject, will have to handle a large variety of data types and subjects
- ongoing national depositories will have to integrate paper archives and electronic archives, assuming that it is unlikely that all paper archives will be converted to digital
- funding has generally been inadequate to support current paper archive activities, let alone the electronic input
- ensuring compliance
- issues of the depositing of digital information have not been worked out in most countries
- there are significant concerns about the security of intellectual property rights by owners if legal deposition is extended to digital works that are then made available on a network

Interoperable Archives: Open Archival Information System Model

The need for increased heterogeneity among archives has led the Consultative Committee for Space Data Systems of ISO to develop an Open Archival Information System (OAIS) architecture (ftp://nssdc.gsfc.nasa.gov/pub/sfdu/isoas/int07/CCSDC-650.0-W-4.pdf.) The goal of this architecture is to provide a framework in which a variety of archives can be developed. It identifies the characteristics of an archive that must be met in order to preserve and provide ongoing access.

The key components of the archive include a series of object-oriented metadata packets that provide the information necessary to ingest, manage, and reuse the archived information. The conceptual data model begins with an Information Package, a conceptual container including two types of information, Content Information and Preservation Description Information. The Content Information includes the actual digital object (i.e., the bits) and the associated Representation Information needed to understand the bits. Descriptive Information is held outside the Information Package and contains metadata necessary for discovery of the resource. The Preservation Description Information portion of the Information Package includes information that is needed to understand the Content Information for long-term preservation. It includes Provenance, which describes the source and history of the Content Information; Context, which describes how the Content Information relates to other Content Information; Reference, which provides one or more identifiers by which the Content Information can be uniquely identified (e.g., an ISBN, Digital Object Identifier, URN, etc.); and Fixity, a wrapper which protects the object against undocumented alteration (e.g., a checksum). In addition to the Information Package data model, the OAIS describes models for Submission Packets, the information needed when an Information Package is submitted by a producer to the archive (or from one archive to another). As the object moves from one archive to another, the Submission Packets are cumulated into an Archival Information Package for preservation. At the access end, a Dissemination Information Package has been defined which provides necessary Packaging Information to allow the user to distinguish the package from others without "opening" the contents of the Information Package itself.

In addition to the conceptual model, the OAIS Reference Model describes the responsibilities and interactions among the entities involved in the archiving process -- producer, archive, management, and consumer. It also includes detailed functional models for the primary activities performed in archiving, such as ingest, archival storage, data management, administration and access.

While the goal of the CCSDS was interoperability of space communication information, they really developed the model with a broader view in mind. After the fact, many on the committee believe that this reference model can be used to develop similar archives for other data types. (Sawyer, Personal communication. 1999.)

The OAIS reference model is being used as the basis for the infrastructure architecture under development by the CEDARS Project (www.leeds.ac.uk/cedars/). The CEDARS project deals with more traditional electronic library materials, but they are designing their network of archives as OAIS compliant archives. While the project is still in its design stage and there are no results to report, CEDARS is "grappling with elements of an implementation - particularly for fitting published material for which there are both rights negotiations and access control mechanisms [necessary]." In some cases the "openness" of the OAIS model is both its strength and its weakness. (Russell, Personal communication, 1999.)

The OAIS may be a generalizable reference model. The concepts behind the model have parallels in other communities. Through such an architecture, it may be possible to not only provide search systems that operate across heterogeneous database structure more easily, but across heterogeneous data types -- numeric, text, video, image and even multimedia. The architecture is presented in an object-oriented fashion which transitions well to XML and RDF applications. However, the language of the current white paper definitely places the model within the data community. In order to make it more widely usable, it will be necessary to interpret the model for various archival communities and stakeholders.

Life Cycle Players and Their Roles

In discussing the various models we touched on many of the stakeholders in the information life cycle. The following sections take a different view of developments and focus on the changing directions and roles of the players themselves.

Creators/Producers

To date there has been little involvement on the part of the creator/producer in the archiving process. In the PANDORA, NERC, DAAC and NDAD archives, the creators/producers may be minimally involved in the archiving process by providing easier access to their materials, by providing documentation, and by signing agreements allowing for various degrees of access. PANDORA indicated that while most of their materials are obtained through harvesting robots, there are cases where a creator/producer "pushes" the material to the archive via ftp or CD-ROM.

However, despite limited involvement to date on the part of creators, it is with the creator that archiving must start in order to be successful and efficient. Compliance on the part of the creator is particularly important in situations such as depositories and corporate archives where either through regulation, coercion or rewards, creators produce digital objects to certain standards. These standards, while they may still be de facto, provide some semblance of order in the chaos. They reduce the number of possibilities that the archive must deal with.

Publishers

Many of the large publishers are creating their own archives based on the provision of their products in electronic form, as counterparts to their print products. Several years ago, publishers began to realize that if the archival version of their efforts were stored in such a way that it could be reused, there would be additional benefit and revenue to be gained from the repository.

For the publisher this has meant attention to not only the technical details of controlling an e-journal, but the intellectual property issues as well. It is important for publishers to consider the copyright agreements that they have with authors, to ensure that they have the right to continued reuse and to migrating the content to new platforms and formats. (Meyers & Beebe, 1998.)

Two of the most notable publishers who have taken on archiving responsibilities directly are the American Institute of Physics (AIP) and the American Chemical Society (ACS). Both of these are learned, professional societies that are charged by their members with preservation of the knowledge produced in their respective fields. They also have long histories as secondary publishers, accustomed to the online environment and the necessity of archiving secondary information for reuse and periodic reloading.

AIP currently archives all journals available via its Online Journal Publishing Service (OJPS). In addition, the supplementary materials are archived in the EPAPS (Electronic Physics Auxiliary Publication Service). In addition to archiving its own publications, AIP will provide archiving as a service to member societies for which AIP provides publishing support, but the cost of such a service has not yet been determined (Ingoldsby, Personal communication, 1999).

AIP is most notable for its well developed Archiving and Usage Policy (http://www.aip.org/journals/archive/index.html). Policies covered in the document, which was developed by a task group of AIP publications staff, librarians, and users, include access rights of current subscribers, lapsed subscribers, and non-subscribers; downloading, and the availablity of physical copies of the archive and the cost for these copies. The policy also identifies to the user AIP's approach commitment to archiving and its approach to technology migration, refreshing of media, and retractions and corrections. In the latter case, the original articles are not altered, but annnotations are made to text explaining the retraction or correction to the article. The AIP has planned for one or more secondary archive sites, which will provide backup and may be used to spread the access across multiple geographic locations. In addition, the primary archive is never used for its online searching service, but is archived to ensure that its contents are not altered.

ACS has 26 journals available via ACS Web Editions (www.pubs.acs.org). While most of the journals are those without advertisements due to concerns over the presentation of the advertisements in an online environment, *Chemical and Engineering News,* which includes advertisement, was recently made available via Web Editions. With the exception of newer journal titles, all archives include material back to January 1996. The ACS is committed to archiving its journals into the future, but it isn't clear whether the back issues that exist only in paper will be scanned and included in this archive (Garson, Personal communication, 1999).

There are unique aspects to the learned societies that make them stand apart as models for digital archiving. The AIP noted that its use of the word "archival journal" to characterize its journals in physics and related subjects, refers to "the longstanding requirement by scholars for a body of literature that reliably records all published and established knowledge" (Scott, Personal communication. 1999.) The electronic archive is seen as an extension of this objective in the electronic era. Unlike commercial publishers, the society publishers believe that they have a mandate from their membership to preserve their publications and continue to provide access, regardless of whether the economics are beneficial.

As an extension of the learned society commitment to preserve the intellectual efforts in its discipline, the American Astrophysical Society collaborates with other astronomical societies to make available a worldwide body of astronomical literature. Each producer maintains its own archive, and links are made and retained between items through references in the full text, cited references, and through links to bibliographic databases in the astronomical sciences. In this case, AAS, identified the possible negative economic impact of this level of archiving and established an escrow account that would take a small amount from current income to migrate to a new system and SGML structure every five years. However, AAS now believes that this will not be necessary, and that these costs for maintenance of access and of the working journal are covered well enough under current operations. (Boyce, Personal communication. 1999.)

Commercial publishers, on the other hand, may be more driven by the economic benefits of the archive for reuse. It is not clear how the economics impact the role of the commercial publisher, since no commercial publishers responded to the survey even though they were contacted. It should be noted that the NEDLIB, networked depository library project funded by the European Union, has sponsorship from Kluwer, Elsevier and Springer. These organizations have also been included in major digital library projects, such as the Los Alamos National Laboratory's "Library Without Walls" and the OhioLINK consortium. Many commercial scientific, technical and medical publishers under the auspices of the STM Publishers group recently held a workshop in Washington, D.C. on April 22-23 on electronic publishing, including DEA issues (209.41.0.61/stm/index.html).

Secondary Services

An informal survey conducted by the National Federation of Abstracting and Information Services (NFAIS) in July 1998 indicated that there was little interest in this topic among the NFAIS

membership of largely secondary publishers (Kaser, Personal communication, 1999). However, this may be more a matter of perspective than reality. Secondary services, particularly in the sciences, have historically considered the longevity of their services as a key asset.

Secondary services have historically been involved in paper archiving in an indirect, but often critical way. For some types of literature, particularly the journal article level, secondary services have served as the initial catalog for discovering that the item exists and may be of interest to the user. Secondary services connected to government agencies and learned societies have had a direct connection to the archive, by providing the location and ordering information for the actual document in the archive. Other commercial and not-for-profit secondary services have had looser connections to the archives through arrangements with large journal archives at research libraries which can serve as document delivery sources for the contents of the secondary catalog.

This traditional role of the secondary service may be reinvented in the electronic archive environment. For example, the National Library of Australia has recently begun efforts to engage secondary services in the provision of metadata that is appropriate for use by the PANDORA catalogs. Provision of metadata that can be harvested and connected to the Australiana that is on the Web, with conversion or enhancement to full MARC cataloging would reduce the intense resource requirements faced by NLA as they proceed with PANDORA.

There has also been some suggestion that the secondary publisher could fill the niche for DEA in certain disciplines (Kelly, 1997). First, the secondary publishers tend to be organized by discipline, aggregating the works of multiple publishers. Unlike primary publishers, they have tended to expand out to non-textual material with many now cataloging Web sites, CD-ROMs and other digital objects.

They may also provide the more extensive finding aids that are necessary to bring together a network of archives across disciplines and for interdisciplinary purposes, because they provide more extensive access points than simple Tables of Content. The secondary arms of publishers such as Elsevier, the American Chemical Society and the American Institute of Physics are becoming increasingly interrelated with the primary arms. In a single publisher environment, they are providing the metadata support that PANDORA envisions. OCLC's Electronic Journal Service is already linking the bibliographic databases available through its ECO and FirstSearch systems to the archived electronic journals. As archives of this nature grow, it is likely that support for the archive provided by the secondary services will also grow.

Libraries and Library Consortia

Much of the concern about the archiving of digital information has come from the library community in relation to the burgeoning development of electronic journals, either with our without print counterparts. Unlike the print model, where a subscription purchased the physical item which the library could archive, the initial electronic journals did not consider archiving at all. Librarians, interested in providing electronic resources for the ease of use of their constituents,

suddenly found themselves paying the same, or more, for the online electronic version with no ownership of tangible goods and no guarantee of access should they be unable to subscribe in the future or should the publisher or the journal cease to exist. Since scholarship and the library role considers this a major responsibility, a number of initiatives were begun.

This situation, along with attempts to reduce the burden of one-off agreements between publishers and libraries, has led to model licenses that begin to address DEA issues. The model licenses and guidelines from the Association of Research Libraries (www.arl.org/scomm/licensing/principles.html), the International Coalition of Library Consortia (ICOLC) (www.library.yale.edu/consortia/statement.html) , and the UK Universities and Publishers (www.ukoln.ac.uk/services/elib/papers/pa/licence/Pajisc21.html) provide standard language related to the archiving of the material received electronically. The ICOLC guidelines suggest that an archival copy shall be provided to the library. The JISC agreement with the Publishers Association in the UK requires that the publisher provide for the archiving either itself or through a third-party repository. In most cases, access to these archives is limited to the members of the organization or to those who can access the archive from a specific geographical location/site.

OhioLink is a consortia of various types of libraries within Ohio (www.ohiolink.edu) . OhioLink was an early advocate of electronic dissemination and sought to provide access to a variety of electronic journals. The Electronic Journal Center (EJC) is OhioLINK's self-operated, multi-publisher, aggregated collection of electronic journals. After analysis of the options, OhioLINK determined that its own site would give them "the best combination of performance, functionality, and integration with other resources and the archive." (Sanville, Personal communication, 1999.) Typically, the complete electronic journal collection of a publisher is licensed. OhioLINK currently has the collections of Elsevier Science, Academic Press, and Project Muse loaded. Upcoming loads will include the American Physical Society (APS), Kluwer, Wiley and Springer. OhioLINK receives (CD-ROM or ftp) and loads bibliographic, table of contents, article abstracts, and article full text data from the publishers. (For APS, they will be loading only the bibliographic record and then linking to the full text on the APS site.) The archive is available online to students, faculty and staff at Ohio higher education institutions. There is no cost to users for access or downloading of articles. However, if a user wants to print an article the local library may have a small per page printing fee. There are other examples of libraries that require that they be allowed to maintain copies for archival purposes as part of their agreements with publishers (University of Michigan, Royal Institute of Technology Library in Sweden).

However, it is unlikely that all major libraries, let alone medium-sized libraries, will be able to take the approach of OhioLINK. The economics of libraries and library consortia are different from other models. Because the digital library concept is based on access rather than ownership, and libraries generally do not have the resources to support large data centers, the responsibility for the archiving of digital objects in this model is likely to fall outside the digital library organization. Even within the large digital library organizations, the library function is often separate from the data center, where the organization's current data may be stored, and also from the archive which

may be hosting relevant organizational electronic records.

The digital library seeks to bring these all together, but the stewardship and ownership are elsewhere. The role of most libraries will be to advocate and where necessary require that DEA issues be covered in licenses and contracts. They may also seek new relationships with third party and institutional archives to achieve this goal. There will continue to be power in consortia and ever changing digital library organizations. "...to realize digital economies of scale can and almost certainly will result in digital libraries that effectively manage their collections by allocating functional responsibilities for their operation largely outside their organization, in ways that are quite different from how we are presently accustomed to seeing them. Indeed, if we look closely at the research university, we can see that the political, economic, and other conditions that shape the use of digital information in this community of our common interest are giving rise, before our eyes, to new and distinctive kinds of library organizations." (Waters, 1998).

Funding Agencies

No funding agencies were interviewed for this report. However, many experts interviewed indicated that successful archives are dependent on both initial and continuing support from funding sources. The most continuous archives in the data and print arenas have had funded mandates. This is still critical in the new environment.

The officials from some of the more established data centers, when asked about the cost issue and what could be done to improve the effectiveness of DEA, indicated that programs conducting research must also fund the appropriate data management. One of the products of such research should be the data project. Among many programs where data management is key, there has been thought given by the information managers to preservation of the data beyond the length of the program. One program has committed to manage the data for the length of the program, but official guidance has not been given. It is now in the sixth of seven years of data collection, with another 6-7 years expected after that. The Management Team estimates that there will be active analysis on the data for at least another 10 years after collection ceases, and ongoing reuse may extend well beyond that. There has been discussion on how to get money to cover data once the program is completed, but there is no commitment yet from funding sources to do so. Much depends on governmental appropriations, which is often a roller coaster process. The one long term solution, that of a data endowment runs quite contrary to the short term cycles of many governments. This is discussed further under the Economics section.

Users

To date, there appears to be little involvement on the part of the end user in the archiving process other than to support the use and, sometimes the funding, of the archives. This is particularly true in the case of the data archives which are not acting as national depositories. Their continued funding depends on the testimonies of the users, and a high degree of customer service is a strategy to keep the archive visible, usable, and funded.

Best Practices by Life Cycle Function

Although DEA is new and the field is complex and changing, as a result of the review of operational and prototype DEA projects, we have identified a number of best practice areas and some examples of active approaches. The section should be of interest for both the operational areas that have emerged, as well as the active current examples of ways to address these areas. The best practices are organized by the applicable stage in the life cycle management -- creation, acquisition/collection development, cataloging and identification, storage, preservation and long-term access.

Creation

All groups involved acknowledge that creation is where the long-term archiving and preservation must start. First, consideration to the long-term value of the information on the part of the creator may be a good indication of the value placed on it by people within the same discipline or area of research in the future. The US Department of Agriculture's Digital Publications Preservation Steering Committee has discussed the concept of having the creator provide a *preservation indicator* in the document. This would not take the place of formal retention schedules, but it would provide an indication of the long-term value that the creator, as a practicing researcher, attaches to the document's contents.

Secondly, the preservation and archiving process is made more efficient when attention is paid to issues of consistency, format, standardization and metadata description in the very beginning of the information life cycle. The Oak Ridge National Laboratory (Tennessee, USA) recently announced guidelines for the creation of digital documents. Limits are placed not only on the software that can be used, but on the format and layout of the documents.

Others in the information creation chain for formal published materials, such as publishers, funding sources, learned societies, etc. can play a large part in promoting such attention on the part of the creators. Governments and institutions are beginning to require a more limited number of formats and attached metadata for objects created under their auspices. As standards groups and vendors move to the incorporation of Extensible Mark-up Language (XML) and RDF (Resource Description Framework) architectures in their word processing and database products the creation of metadata as part of the origination of the object will be easier. However, work remains to identify the specific data elements needed for long-term preservation as opposed to discovery, particularly for non-textual data types likes images, video and multimedia.

Acquisition and Collection Development

The most extensive acquisition and collection policies have been developed by the national libraries involved in digital archiving. This is primarily because there continues, in most cases, to be questions about legal deposit of digital materials, and guidelines are helpful to establish the boundaries. As the NLC notes in its recently published collection guidelines, 'The main difficulty

in extending legal deposit to network publishing is that legal deposit is a relatively indiscriminate acquisition mechanism that aims at comprehensiveness. In the network environment, any individual with access to the Internet can be a publisher and the network publishing process does not always provide the initial screening and selection at the manuscript stage on which libraries have traditionally relied in the print environment. In addition, because electronic publishing is innovative and changing in nature, legal deposit legislation should remain open-ended enough to incorporate a wide range of existing and potential electronic materials and should stipulate as few restrictions as possible. Selection policies are therefore needed to ensure the collection of publications of lasting cultural and research value." (www.nlc-bnc.ca/pubs/irm/eneppg.htm)

Similarly, even though the goal of the PANDORA project is the preservation of Australian Internet publishing, it is impossible to archive everything. Therefore, the NLA has formulated Guidelines for the *Selection of Online Australian Publications Intended for Preservation by the National Library of Australia* (www.nla.gov/au/scoap/guidelines.html). These guidelines are key to successful networking of the state libraries into the National Collection of Australian Electronic Publications.

Scholarly publications of national significance and those of current and long term research value are archived comprehensively. Other items are archived on a selective basis "to provide a broad cultural snapshot of how Australians are using the Internet to disseminate information, express opinions, lobby, and publish their creative work." In all cases, NLA, in the absence of digital deposit legislation, seeks permission from the copyright owner before copying the resource for the archive. (Phillips, Personal communication, 1999).

The major document types archived by PANDORA include:

- Monographs - fixed content as in a traditional print publication cumulative or evolving, whose contents change over time

- Serials

  - regular serials - issues appear sequentially in traditional print publication patterns, and fit the definition of 'serial' for cataloguing purposes and for assignment of International Standard Serial Numbers (ISSN)

  -evolving serials, whose contents change over time

- Home pages

- Ephemera (the Guidelines include an entire appendix dedicated to the selection of ephemera)

The specific criteria for selection include:

- a significant proportion of Australian content or be on a subject of social, political, cultural, religious, scientific or economic significance and relevance to Australia and be written by an Australian author

- the sole version of a work, or, if the work has multiple versions such as print or microform in addition to the online, the online has significant additional information or value.

.

- Authority and long term research value (Support or sponsorship by an official funding body would be one factor only, which might influence a decision in favor of selection.)

- Topical issues as determined by the Collection Development Manager

Although, **content is the pre-eminent factor determining selection,** selection is also based on the ability of the archive to successfully handle the digital object technically. Sometimes there are pages that depend on programs that reside on the publisher's server, such as pages that are created "on-the-fly." PANDORA has not successfully archived these types of pages to-date. (Phillips, Personal communication, 1999.)

The Royal Library, National Library of Sweden takes an entirely different approach to collection development (kulturaw3,kb.se/html/projectdescription.html). Instead of evaluating and selecting material, the Kulturaw3 approach is to run a robot periodically to capture sites from the .se domain and from known Web servers that are located in Sweden even though they have .com extensions. In addition, some material is obtained from foreign sites with material about Sweden, such as travel information or translations of Swedish literature. The Swedish opinion is that it is impossible to know now what will be of value in the future, so they are not making value judgements. However, they have set priorities for periodicals, static documents, and dynamic documents such as HTML pages. Conferences, usenet groups, ftp archives, and databases are considered lower priority. In the most recent reported run of the robot in late 1997, the robot found 9.5 million URLs from 26,000 Web sites. Of these, about two-thirds were found based on the .se extension.

The EVA Project at the Finnish National Library uses techniques similar to those used in Sweden. However, the guidelines from EVA identify issues to be considered when harvesting using robots. In order not to overload the servers being harvested, particularly the public networks, EVA has established time limits. Time limits are set to about 60 seconds between visits to a single Web server and 1 month between capturing and recapturing a single URL. Developers at EVA consider this approach to be "very rough and not flexible enough for archiving purposes" (Helsinki University Library and Center for Scientific Computing in Finland, No Date.), preferring that the time limits be more configurable at the server and preferrably at the individual URL levels. In practice this means that the scheduler must be a database application which can be modified by the librarian.

For data centers, dataset content is determined by expert reviewers, some internal and some based

on external peer advisory groups. These centers are very mature in terms of their content evaluation functions. When resources are not sufficient to cover maintenance of all datasets, then this peer review determines what shall continue to be preserved.

Determining Extent

Connected to selection is the issue of extent. What is the extent, the boundaries, of the digital work? How high or low within the work do you archive? For publishers and repository agents this is not a question, because the extent is determined by the originator. However, this is not the case for the national libraries and depositories. They must establish guidelines for extent. As the NLA Guidelines state, "if a publication has a number of internal or external links, the boundaries of the publication need to be decided. [For PANDORA] Internal links only are archived. Both higher and lower links on the site are explored to establish which components form a title that stands on its own for the purposes of preservation and cataloguing. [For PANDORA,] preference is given to breaking down large sites into component titles and selecting those which meet the guidelines. However, sometimes the components of larger publications or sites do not stand well on their own but together do form a valuable source of information. In this case, if it fits the guidelines, the site should be selected for archiving as an entity." The Web harvester used by PANDORA is "programmed" to select only those URLs that are in the same directory or in subdirectories of the URL that is provided. Similar guidelines are used by the EVA project in Finland.

Archiving Related Links

An interesting issue raised by the hypertext linking of digital objects, is the question of what should be archived? Is the object the single source item or its related hypertext links? What about a document that is made up of a series of links, connected by a Table of Contents page? What about citations and references that are links? This issue has been addressed by projects in a variety of ways.

Most organizations archive the links, but not the text of the linked cites. AIP archives the links (URLs or other identifiers) but not the text or content of any of those links, unless the linked item happens to be in its publications archive or in the supplemental material which it also archives. Similarly, DOE OSTI does not intentionally archive any links beyond the extent of the digital object identified. However, the document may be linked to another document if that document is another DOE document in the OSTI archive.

In a slightly different approach, the National Library of Australia has chosen to archive the linked item, only if it is on the same server as the source item it is archiving, believing that there is less likelihood that the hypertext-linked item will disappear, unless the original source does as well. Similarly, the Electronic Publications Preservation Project determined that "After the difficulties involved in tracking down hypertext links and acquiring the linked objects were considered, a hypertext electronic publication was defined as consisting only of linked objects stored on one

Internet domain. The previous issue of the same periodical, accessed through a hypertext link, would be considered a part of the original publication. Another publication accessed through a hypertext link would not be considered part of the original publication, because it is impossible for the NLC to maintain or preserve the integrity of links to other publications or Internet domains." The EPPP proposed that the hypertext links only to the first level be archived. At OhioLink's Electronic Journal Center, links to other resources are only supplied once the full data is loaded locally. The linked content is not archived, but exists only "on the fly" when the user selects the link. The viability of the links is not tested during the loading process. The Internet Archive of Brewster Kahle, is, of course, archiving all links (unless they are to "off limits" sites), because its aim is to archive the entire Internet.

The international system for astronomical literature, on the other hand, maintains all links, to both documents and supporting materials in other formats, based on extensive collaboration among the various astronomical societies, researchers, universities, and government agencies. (Boyce, Personal communication. 1999.) Each organization archives its own publications, but links are maintained not only from references in the full text and cited references of the articles, but between and among the major international astronomical databases.

Cataloging and Identification

Both cataloging and identification allow the archiving organization to manage the collection. Cataloging in the form of metadata provides supports organization, access and administration information. Identification provides a unique key for finding the object itself and linking that object to other related objects. Cataloging and identification practices are often related to what is being archived and the resources available for managing the archive.

Metadata

Some form of metadata is used for all archives. Metadata exists for description, reuse, administration, and preservation of the archived object. The level at which metadata is applied depends on the type of data and the future purposes to which it may be put. Datasets are generally cataloged at the file or collection level. Journal articles are cataloged individually, sometimes with no concern about metadata for the issue level. Homepages provide a particularly difficult problem for determining the level at which metadata should be applied. Generally, the metadata is applied to whatever is considered to be the full extent of the resource.

In general, the metadata files are stored separate from the archives themselves. Libraries may store the metadata in their online public access catalogs. Publishers may store the metadata in a bibliographic database. However, in some instances, such as electronic journals with SGML headers, the information may be stored in the archive itself and extracted for the catalog. In the case of distributed archives, the metadata may be stored centrally with the electronic resources distributed. Depending on the search tools used, the metadata may be stored as embedded tags in the online resource.

A variety of metadata formats are used, depending on the data type, discipline and cataloging resources available and approaches used. MARC cataloging is used by the national libraries, with some fields unable to be filled and others, such as the 856 taking on new meaning, as it contains the URL or the Digital Object Identifier. MARC is used by NLA, NLC and the EVA project. NLA also uses a Dublin Core-like format in cases where this supports receiving metadata from the publisher, eliminating the library cataloging. In the U.S., the IAW DoD Standard 5015.2 and the National Image Mapping Agency Core Video Metadata Profile serve as the basis for the multimedia metadata for the DITT Project.

The attributes and the metadata content considered of interest when describing a particular object vary based on data type, origin, future use, and discipline. For example, attributes of the NASA DAAC metadata include the instrument generating the data, the date and time, other existing conditions, quality factors, etc. Part of the Defense Information Technology Testbed Project has involved identifying core and unique elements.

Discussions surrounding the interoperability of archives, both within and across disciplines, focus on the need to be able to cross-walk the various metadata formats. This is key to the ability to network heterogeneous data types and disciplines. The OAIS Reference Model considers this issue by encapsulating specific metadata in a consistent data model. The LTER has developed mechanisms for "fitting" its network-specific metadata information into the broader scheme of the Federal Geographic Data Committee and other emerging standards related to the discipline of ecology and related sciences.

The creation of metadata differs substantially depending on the type and volume of the original data object. For data centers, much of the data is "created" by the measurement or monitoring instruments themselves, and the metadata is supplied along with the data stream. This may include location, instrument type, and other quality indicators concerning the context of the measurement. In some cases, this may be supplemented by information provided by the original researcher. For smaller datasets and many "publications" much of the metadata continues to be created "by hand".

However, across the DEA stakeholders there is continuing interest in automatic generation of metadata, since this is often considered to be a major impediment to archiving more digital electronic information. A project is underway at the U.S. EPA to derive metadata at the data element level from legacy data collections (Shepanek, Personal communication. 1999.). The DITT Project is also investigating fully automated metadata generation.

Ensuring Persistence through Identification

For those archives that do not copy the digital material immediately into the archive, the movement of material from server to server or from directory structure to directory structure on the network, necessitating a change in the URL, is problematic. The use of the exact server as the location identifier both for the source work and any linked works results in lack of persistence

over time. While it is not the intent of this report to describe all the research and projects in this area, this is an area of concern for archives.

Despite possible problems, most archives continue to use the URL when referencing the location for the digital object. In the case of libraries, this is often entered as the content of the 856 field in a standard MARC catalog record. The OCLC archive uses PURLs (purl.oclc.org/), persistent identifiers to which the changeable URL is mapped. ACS uses the Digital Object Identifier, and also maintains the original Manuscript Number assigned to the item at the beginning of the publication process.

A more extensive identification system is used by the AAS. Name resolution is used rather than storing the URLs. In addition, the AAS uses astronomy's standard identifier, called a "Bibcode", which has been in use for fifteen years (Boyce, Personal communication. 1999.) In the Spring of 1999, AAS will add PubRef numbers (a linkage mechanisms originally developed by the National Library of Medicine), and other identifiers can also be added as needed to maintain links.

The Digital Object Identifier (www.doi.org/) is a scheme for persistent identification of a digital object. The DOI Foundation has developed the standardized structure, based on the "handle" technology developed by CNRI. To support the resolution of these DOI's to the actual server location of the item, there needs to be a DOI resolver database. Efforts are now underway within the Foundation to identify the elements that should be present in the database, including those that will be needed for long-term intellectual rights management. (www.doi.org/policy.html) A core set of elements have been defined, with extensions possible for specific genres, such as journal articles. A draft paper on how DOI's can be used as reference links is also available. Because many of the members of the DOI Foundation are commercial publishers, there is a focus on rights management issues in their efforts. However, since this is also a key factor affecting digital electronic archiving, there is much to learn from the demonstration projects planned by the DOI and projects such as NEDLIB which are using the DOI. Attention should also be paid in the development of the metadata that accompanies the DOI and other schemes to elements needed to allow recreation of the "look and feel" and ensure format access.

Storage

Storage practices relate to the plans for migrating from current hardware and software environments to newer environments, the refreshing of media, and backup and recovery.

Hardware/Software Migration

One of the issues that makes digital archiving more urgent than the archiving of traditional formats such as paper is the speed with which technologies are changing. New releases of databases, spreadsheets, and wordprocessors can be expected at least every two-three years, with patches and minor updates more often. While software vendors generally provide migration strategies or upward compatibility for some generations of their products, this may not be true

beyond one or two generations. This is not guaranteed to work for all data types and becomes particularly questionable if the information product has used sophisticated features of the software. There is generally no backward compatibility, and if it is possible, there is certainly loss of integrity of the product.

In addition to software, the hardware landscape is changing almost as rapidly. Storage media have changed, with legacy information perhaps lost forever on older magnetic tapes. Block sizes, tape sizes, tape drive mechanisms and operating systems have changed over time. The movement is particularly evident in the consumer market where 8-track tapes, gave way to audio cassettes, and then CD's and DVDs. There is no easy way to migrate; a digital master is generally required to replicate the quality of the original work.

The most common solution to this problem at this point is migration. This involves keeping up with the software and migrating to new hardware frequently. This is expensive and there is always concern about the loss of data or problems with the quality when a transfer is made. Check algorithms are extremely important when this approach is used. Data centers have been acutely aware of this issue for years with some opting for ASCII as a more universal format.

All the archives queried in the survey were considering the migration issue. However, most of them had not been in existence long enough to face this problem. A common answer was that they would move to the most appropriate technology when needed.

Among those who had considered migration, the migration from one storage media to another was most commonly discussed. Most organizations that responded to the question about the periodicity of media migration, indicated a 3-5 year cycle. The most rigorous media migration practices are in place at the data centers. The ARM Center plans to migrate to new technologies every 4-5 years. During each migration, the data is copied to the new technology. Each migration will require 6-12 months. "This is a major effort and may become nearly continuous as the size increases." (McCord, Personal communication. 1999.)

Plans are less rigorous for the migration to new hardware and software. While the cycle for technological change may be longer, the impact is much greater than media migration. One center manager indicated that basic technologies could not be expected to last longer than a decade (Darwell, Personal communication. 1999). In order to guard against major hardware/software migration issues, the organizations try to procure mainstream commercial equipment. For example, both the American Chemical Society and the U.S. EPA have purchased Oracle, not only for its data management capabilities, but because of the company's longevity and ability to impact standards development. Unfortunately, this level of standardization, and ease of migration, is not as readily obtained among specialized fields, such as climatology and meteorology, where specific systems components are required to interface to instrumentation and to handle the volume of data to be stored and manipulated.

An alternative to migration that is being explored is called the Digital Rosetta Stone

(info.wgbh.org/upf/slides/index.html). This calls for the encapsulation of the software with the product upon archiving. It uses metadata that specifies how to recreate the format. For example, a TIFF image might be typed as such and then metadata information provided that indicates how to reconstruct what a TIFF image is at the engineering --- bits and bytes -- level. An alternative to encapsulating this with every instance of the data type is to create a registry that can be referenced more generally. This registry would uniquely identify the hardware and software environments and provide information on how to recreate the environment in order to preserve the use of the digital object. (Heminger & Robertson, 1998) (tuvok.au.af.mil/ au/ database/ research/ ay1996/afit_la/ rober_sb.htm)) This approach has been greatly expanded in a recent report to the CLIR from Jeffrey Rothenberg (www.clir.org/pubs/reports/rothenberg/contents.html).

However, at this point, there is no system in place to provide the extensive documentation and emulation software required for this approach to be operable, particularly to allow an archive to deal with the variety of older technologies in place. However, advances in this area are being watched by several of the archives, including NLA.

The situation for the foreseeable future will be migration, with emulation coming into play as it is supported by hardware and software manufacturers. The best practice at this point is to keep up with the changes in hardware and software and to plan migrations. However, as the ARM Archive at ORNL noted, the migration will become almost a continuous process as the size of the archives grows to terabytes (McCord, Personal communication. 1999).

Refreshing the Media

In addition to the large-scale migrations due to hardware and software changes, digital archives must address the issue of media refreshment. Because no medium exists that does not deteriorate, it is necessary to copy the contents from the physical medium to a new physical medium. Many archives do this on a routine basis as part of their back up and recovery procedures, with most backup copies being put on new physical medium rather than rewriting over old media.

Backup and Recovery

Most archives indicated that industry standard backup and recovery procedures are used for their archival data. This includes periodic backups to magnetic tape or optical disk. A copy is generally held on-site for near-term recovery with long-term off-site storage for disaster recovery. As AIP noted, an offsite copy of the whole archive would provide a recovery channel in the case of loss of or damage to the prime archive. ACS also noted the importance of these routine procedures. All projects contacted performed routine backup and recovery procedures, with most mentioning that backup tapes are stored at remote sites. The NASA Goddard Space Flight Center DAAC noted that it does not currently have complete recovery of all its datasets but this is being addressed (Sawyer, Personal communication. 1999).

It is interesting to note that the tradition to back up electronic media and have off-site disaster support is helpful to the overall goals of archiving electronic media. This has no real direct parallel in the print world.

Preservation

Preservation is the aspect of archival management that preserves the content as well as the look and feel of the digital object. It also includes decisions related to retention and disposition.

Refreshing the Site Contents

In cases where the archiving is taking place while changes or updates may still occur to the digital object as with the archiving of electronic journals, there is a need to consider refreshing the site contents. This is particularly true of the national depositories. For example, NLA allocates a gathering schedule to each "publication" in its automatic harvesting program. The options include on-off, weekly, monthly, quarterly, half yearly, every nine months, or annually. The selection is dependent on the degree of change expected and the overall stability of the site. Obviously, the burden of refreshing the contents increases as the number of sources stored in the archive increases. However, in the NLA procedures, there is no retention of previous versions of the site, once the successful downloading of the new site contents has been verified. This implies that interim versions have no historic value. The case is different for some data centers, where old datasets are preserved in version forms.

Retention

Retention is a major issue for all archives, whether paper or electronic. Even archives that never discard anything have retention policies to this affect. The national archives have formalized retention schedules which are used to determine the deposit of information from the relevant agencies as well as what is retained by the archive itself.

The retention policies are generally imposed by the sponsor. The DAACs are particularly bound by the retention policies imposed by NASA. However, they do not have NARA-like retention guidelines. For the CDIAC, retention is based on best efforts and the judgements of center scientists on an individual dataset basis, rather than by schedules that address types of objects.

A study of the impact of electronic publishing on small society publishers by the UK Online Library Network's (UKOLN) E-Library Programme identifies another issue for publishers. The publishing and access mechanics of electronic publishing may result in a move to "by the drink" access and payment. However, many of the articles currently published will have little current readership and perhaps even less readership as part of an archive. Unfortunately, it is almost impossible to determine at this point (even with a crystal ball) whether within that set of items with low commercial viability resides the seminal article. (Consider the case of Medelev's seminal article on genetics, which was not considered seminal until almost a century after it was written.

For the public good and the benefit of the future of science, there needs to be a mechanism that addresses the retention concerns not only in terms of today's quality and commercial viability, but in what might be of value in the future. This is not new for the digital environment, but its dimensions are increased by the ease and variety of electronic publishing. The UKOLN study suggests that while market forces may prevail for electronic publishing, there may need to be public intervention to "ensure the archiving of all articles which are published, to enable continued access when commercial provision is no longer profitable." (Fishwick, Edwards and Blagden, 1998).

Standards, Transformations vs. Native Formats

One of the paradoxes of the networked environment is that in an environment that is so dynamic and open to change, there is a greater and greater emphasis on standards. Those who have been archiving for a long period of time have indicated that while they started out with a large number of formats --- primarily textual -- the number of current formats has decreased over time . The market forces have reduced the number of major players drastically. DOE began its project with a limited number of acceptable input formats, because the number of native formats were so large. In the political environment of that time, it was difficult to gain support for the standardization of word processing packages. However, documents are currently received in only a few formats. Text is received in SGML (and its relatives HTML and XML), PDF (Normal and Image), WordPerfect and Word. Images are received in TIFF Group 4 and PDF Image.

Consolidation has occurred in the number of spreadsheet and database formats to a lesser extent. However, there is even less consistency in the modeling, simulation and specific purpose software areas; much of this software continues to be specific to the project.

However, with the network environment has come an increased desire to integrate, interact and interoperate. Therefore, the emphasis in these areas appears to be on the development of standards for interoperability and data exchange, realizing that perhaps the market forces will not play as large a role here as with more general purpose formats. There are significant efforts under way, particularly within the Internet Engineering Task Force (IETF) and other groups to provide overarching interoperability for the Web. This is also true at the discipline level, with efforts under way particularly within the geospatial community. Even though GIS has a limited number of vendors, these vendors have been working together on open GIS standards.

Publishers provide several examples of data transformation. AAS and ACS transform the incoming files from LaTex, Word, or WordPerfect to an SGML tagged ASCII file. AAS believes that this transformation, and reinventing the publications process to "think electronic first" as saved money. "The electronic master copy, if done well, is able to serve as the robust electronic archival copy. Such a well-tagged copy can be updated periodically, at very little cost, to take advantage of advances in both technology and standards. The content remains unchanged, but the public electronic version can be updated to remain compatible with the advances in browsers and other access technology." (Boyce, 1997.)

The data community also provides some examples of data transformation. For example, the NASA DAACs transform incoming information into standard CDF format. NDAD transforms the native format into one of its own devising, since they could not find an existing standard that dealt with all their metadata needs. However, the bit-wise copies are retained, so that someone can replicate what the center has done. (Ashley, Personal communication. 1999.) These transformed formats are considered to be the archival versions.

Preserving the Look and Feel

There are several approaches used to save the "look and feel" of the journal article. The majority of the projects reviewed use either image files (TIFF), PDF, or HTML. TIFF is the most prevalent for those organizations that are involved in any way with the conversion of paper backfiles. For example, JSTOR processes everything from paper into TIFF and then OCR's the TIFF image. The OCR, because it cannot achieve 100% accuracy is used only for searching. The TIFF image is the actual delivery method that the user sees. However, this does not allow the links from these journals to other material on the Web to be maintained as actual links.

HTML/SGML is used by many large publishers, following years of converting publication systems from proprietary formats to SGML. (AAS has a richly encoded SGML format that is used as the archival format from which numerous other formats and products are made. However, XML may be considered in the future. (Boyce, Personal communication. 1999.)) HTML is often provided by downgrading the SGML version that is actually stored by the publisher. PDF versions can also be provided via conversion routines.

For purely electronic documents, PDF is most prevalent. This provides a replica of the Postscript format of the journal, but is reliant upon proprietary encoding technologies. PDF is used in cases where the publication process is less formal, for example with gray literature, theses and dissertations. The Royal Institute of Sweden Library transforms dissertations that are received in formats other than PDF to PDF and HTML (Forsberg, Personal communication, 1999). It is also prevalent as a distribution format among more formal publications.

Several years ago there was a major concern about the use of PDF for long-term storage, because it is a proprietary format. However, there appears to be little concern within the publishing community at this time. The main impetus is less likely to be its acceptability as an archival format as that it retains the look and feel of the original, can be produced and read easily by freeware products, and has a variety of tools available at modest costs that allow for full text searching. Hypertext links are also maintained, which is not true of TIFF images. While PDF is increasingly accepted, concerns remain for long-term preservation, particularly within the national archives and libraries. Even though DOE OSTI accepts and disseminates in PDF format, it continues to work with the U.S. National Archives and Records Administration on the acceptance of PDF as a national government depository format (Langford, Personal communication. 1999.).

Access

All of the previous archival life cycle functions are performed for the purpose of ensuring continuous access to the material in the archive. Successful practices must consider changes access mechanisms and rights management requirements.

Access Mechanisms

All projects reviewed had or are planning Web-based interfaces to the data. The Web interfaces may not be accessing the data directly in the case of SGML and ASCII archives. In some cases, the access is provided to databases that can be searched or to HTML files that can be more easily displayed with current browsers. Additional interfaces are available for certain specialized information. For example, there are GUI interfaces available for the NASA DAAC datasets. In some cases, specific software which must be downloaded first is used to access the data.

Many respondents consider the access and display mechanisms as another source of change in the digital environment. Today it is the Web, but there is no way of knowing what it might be tomorrow. One futuristic example is the use of multi-sense virtual reality. In some cases, it may be possible in the future to enhance the quality of presentation of items from the digital archive. NLM's *Profiles in Science* product creates an electronic archive of the photographs, text, videos, etc. that are provided by donors to this project. This electronic archive is used to create new access versions as the access mechanisms change. However, the originals are always retained. "The evolution of technology has shown that whatever level of detail is captured in the conversion process, it will eventually become insufficient. New hardware and software will make it possible to capture and display at higher quality over time. It is always desirable to capture and recapture using the original item (McCray, Personal communication. 1999.).

Intellectual Property and Rights Management

One of the largest access issues involves rights management. What rights does the archive have? What rights do various user groups have? What rights have been retained by the owner? While many of the national libraries and other depositories consider the archiving to be in line with the legal depository requirements, the deposit legislation has not yet been extended to include electronic materials. Therefore, they are seeking licensing agreements with publishers prior to performing the archiving function. Some national libraries (NLC, NLA) are seeking agreements with the rights owners prior to the archiving of the material. Other libraries have so automated their process that this is not done (Finland, Sweden). Third party archives, with the exception of the Kahle's Internet Archive, generally, seek permission or have permission transferred to them by the rights owner. The Internet Archive, however, will seek to copy everything on the Internet. However, sites that are restricted by password, IP address or more sophisticated mechanisms will not be included. It will also not include sites that have specifically requested not to be included.

In addition to the impact of intellectual property on archival collection development, there are format implications. For example, an interesting intellectual property issue for the NLC was that of format preservation. Even with the electronic journals (45 all together) that they focused on in

the pilot study, there were a variety of formats from ASCII to PDF to word processing formats and tagged SGML. There is the major issue of what to do about this, especially as it relates to the long term preservation and access. "According to the Copyright Act, the NLC only infringes on the author's right of integrity if it distorts, mutilates or otherwise modifies the work, or if it associates the work with a product, service, cause or institution, to the prejudice of the author's honour or reputation. After much discussion, it was decided that converting an electronic publication to a standard format to preserve the quality of the original and to ensure long-term access does not infringe on the author's right of integrity."

Security and version control are also issues that have been discussed in relation to digital preservation. Brewster Kahle raises many interesting questions concerning privacy and "stolen information," particularly since the Internet Archive policy is to archive all sites that are linked to one another in one long chain. (Kahle, 1997.) Similarly, there is concern among image archivists that images can be tampered with without detection. Particularly in cases where conservation issues are at stake, it is important to have metadata to manage encryption, watermarks, digital signatures, etc. that can survive despite changes in the format and media on which the digital item is stored.

Practices Related to Specific Formats and Data Types

In addition to the best practices related to specific life cycle functions, best practices have been identified for certain formats and data types, including text, images, and multimedia.

Character Sets for Numeric and Textual Information

The homogeneity of the data archives among the NASA DAACs means that the basic information could be kept in ASCII which has been generally accepted as an archiving standard for text and numeric data, with proper structural encoding and documentation. One of the publishing community's initial goals with the Standard Generalized Markup Language (SGML) was to use ASCII as the standard format – a format which can easily be migrated to different platforms, different database structures, and different software, with the SGML DTD providing the key to understanding the encoding used for page representation. The next likely enhancement in character encoding will be Unicode which extends ASCII to handle all languages and special characters. Since the lowest level of Unicode is standard ASCII, it will migrate very easily. The encoding done to handle special characters could be transferred, but with proper documentation this may not be necessary depending on the way in which the archive is used.

Resolution and Compression Considerations for Images

Some of the most extensive work related to images has been done within the National Digital Library Project at the U.S. Library of Congress. Beginning in the early 1990's, the American Memory Project has digitized over 30 historical collections including digitized documents, photographs, sound, moving pictures and text. While most of them are related to American

government, history, culture and the humanities, there are lessons to be learned from some of these efforts.

In October 1998, the Library of Congress Manuscript Digitization Demonstration Project final report was released (memory.loc.gov/ammem/pictel/index.html). The project made the distinction between "preservation-quality" and "access-quality images". Preservation quality are designed to withstand the test of time and to be of a high enough quality that it can take advantage of new image search, display and storage technologies. "This project is dedicated to the development of specifications for images that, if longevity can be promised, will
serve the goals of preservation, i.e., will serve as reasonable substitutes in the event that the original item is lost or deteriorates." Access-quality image specifications acknowledge that under the current situation, there is generally insufficient band-width to transfer the high-quality (highest resolution) images. "Such images would be lower in either spatial resolution ("dots per inch") or tonal resolution ("bits per pixel") or both, and derived, if possible, from the preservation-quality images. Lower-resolution images--whose digital files should be smaller in extent (bytes)--can be more easily handled in computer systems. The project sought to identify images that, although less faithful to the original than preservation-quality images, offer high legibility and good service to researchers."

The LoC committee discussions indicated that there may be differences in the acceptable quality between manuscripts (which were the focus of LoC's research) and other types of images, such as pictures in monographs or journals which are generally halftone and require high quality for replication.

A major issue related to images is the progression of compression routines. While improved routines are necessary to keep the sizes of the large preservation-quality image databases in check, these compression routines themselves can create difficult migration paths. Several respondents indicated that they considered the changes in compression algorithms to be more detrimental to preservation activities than technology migration.

A recent NISO/RLG/CLIR workshop on Metadata for Image Preservation addressed not only issues related to resolution and compression, but the whole range of image description and metadata elements needed to understand the image technically. There was significant discussion of issues related to preservation, particularly the movement of an image from one collection to another, on its way to longer-term permanence, and how to verify the authenticity of the image over time. The metadata being developed is intended to bridge the resolution and compression issues over time. As draft elements, guidelines, and white papers are developed, they will be made available on the work groups Web site (www.niso.org/images.html).

Object Archiving for Multimedia

All the issues related to the various data types, and more, are bundled into the issues surrounding the archiving of multimedia works. Since efficient archiving, access, reuse and preservation differ

based on data type, multimedia, which combines various data types, cannot be dealt with by a single approach.

The Defense Information Technology Testbed (DITT) Project between the U.S. DoD and the Joint Analysis Center, UK is developing a long-term archive of multimedia objects in support of military operations and to feed training and lessons learned systems. (Borkowski, Personal communication, 1999.) Multimedia is particularly important in the U.S. military's distance learning efforts, as well as its strategic "Army After Next." The prototype was demonstrated in Bosnia. The multimedia archive will be used to create training and doctrinal materials under the Advanced Distance Learning Initiative. Another long term goal is to preserve imagery that lead up to important battle decisions. This prototype is a "proof of concept" for full-life management of multimedia records within the operational environment and as a first step toward a virtual research library for Unmanned Aerial Vehicle (UAV) video in particular and multimedia imagery in general. The prototype showed that these components could be collected, linked, searched and managed as "one record."

UAV records include MPEG, JPEG, audio, text and metadata components. VHS video is recorded by reconnaissance UAVs. From the command post, a remote pilot narrates the mission in audio. The video is converted to digital MPEG and audio is digitized (wav). The video and audio are encrypted and transmitted real-time to the JAC in Molesworth, England. There the audio file is automatically transcribed to text, metadata is generated, a mosaic file (a JPEG file that summaries the overall track) is automatically generated, and the mission profile file is extracted from operator entered mission data. On a 30-day schedule, the magnetic tapes containing this information are sent to Ft. Leavenworth, Kansas, US where they are loaded into the MAAS system, a multimedia data warehouse. The MAAS can be searched for specific video clips, images, etc. DoD is also experimenting with the use of the RetrievalWare search engine for retrieval of video data based on content-based querying.

In addition to the issues of archiving each object that makes up the multimedia collection in a way that is most appropriate for each data and object type, it is necessary to bring the collective multimedia object back together for reuse. The San Diego Supercomputer Center using a "container" architecture, with a hierarchical storage architecture and a special Resource Broker System to store and retrieve metadata and objects in collections. Each level of the collection hierarchy has appropriate metadata required for preservation, reuse, and reassembly of the collection. The metadata allows the system to reconstruct the organization of the collection based on the individual disparate objects. This structure has been used in pilot projects with patent data for the US Patent and Trademark Office and with various administrative information formats from the U.S. National Archives and Records Administration. SDSC is also involved in pilot projects with ecological data (with NCEAS and LTER) and art images (the RLG AMICO project).

The major complexity with multimedia is not only the combination of various data types and their interaction, but the fact that much of the "look and feel" of multimedia is dependent on the software under which it runs. It is often difficult to separate multimedia from its software and

hardware environments. Multimedia's reliance on hardware and software environments emphasizes the problem of how to reconstruct these "systems" in the future. Jeffrey Rothenberg of the Rand Corporation has espoused the development of emulation capabilities that can replicate the "behavior" of the system. By storing and archiving complete definitions of the hardware and software's behavior, a specific piece of software running on a Pentium could be replicated in the future. The proprietary nature of much of these definitions is an issue. The possibility of requiring deposition of these definitions in a software/hardware registry/depository has been discussed.

An alternative approach which may support the archiving of multimedia in the future is the Advanced Authoring Facility which is a Microsoft-backed industry initiative to specify an extensible, platform-independent multimedia file format (www.microsoft.com/aaf/). Industry participants include Adobe, Avid, and Pinnacle. While the main purpose for the current work is to support authoring interchange by AAF-compliant multimedia content creation tools, this has ramifications for preservation, particularly when it is integrated with efforts such as the Universal Preservation Format (info.wgbh.org/upf/).

Practices Related to Specific Object (Document) Types

There are also specific practices related to certain objects or document types, such as biological sequence data, software and datasets.

Biological Sequence Data

Biological sequence data banks have some unique aspects that must be taken into account when dealing with digital archiving. While these data banks have only gigabytes of data, rather than the terabytes archived at data centers such as the NASA DAACs, the data has a more encoded structure. This results in the need for extensive validation routines to ensure the quality of the information when it is submitted by the researcher and as the information is migrated. (Benson, Personal communication. 1999.) NCBI has approximately 30 Ph.D.s who act as quality assurance specialists, reviewing the information manually, even after it has passed through a variety of validation algorithms. The ongoing nature of research into the DNA sequences also results in corrections and additions to a particular sequence record. A history of changes must be maintained. All changes are controlled by NCBI, with approval by the sequence owner.

In addition, the need for validation, searching and reporting means that a database is important to the continuing use of these data banks as active archives. The Protein Data Bank, formerly maintained by the Brookhaven National Laboratory, is now being transitioned to the Collaboratory for Structured Bioinformatics a non-profit consortia (Rutgers University, the San Diego Supercomputer Center, and the National Institute for Standards and Technology). It is undergoing a major database and system migration (Fagan, Personal communication. 1999.) For archival purposes, the sequence data was always held in simple ASCII files. Unfortunately, this limited its ability to be searched quickly. In the new data structure, the old files will be

maintained, but they will also be provided in a structured Oracle database. The GenBank data bank maintained by the National Center for Biotechnology Information at the U.S. National Library of Medicine has encountered similar issues related to the long-term preservation of DNA sequences. ASCII is used as the preservation and distribution format; a Sybase database provides the structure for searching, reporting and maintenance. (Benson, Personal communication. 1999.)

Documentation for Software and Datasets

While no specific projects were identified in this area, it is of increasing concern to both technologists and scientists in certain disciplines. As the use of computers and computer generated research results grows, many results cannot be verified or reproduced without access to the proper software, much of which may be homegrown. This has led to projects to better document software, particularly for data generation and analysis. Part of the NDAD project involves guidelines for the documentation of datasets, models and supporting software.

The software industry itself has created software library repositories for preservation and reuse. Microsoft is developing vast libraries of software objects that can be reused and recombined to make the most of the intellectual capital investments. The commercial software industry may be a group to which the sciences can look for best practices involving the archiving of software.

Cost/Resources

Although cost is recognized as a basic driver in DEA, it was also the most difficult aspect on which to gather information. In some cases, a lack of response was because of the proprietary nature of this information. However, in most cases, the respondents indicated that they just didn't know how much the archive was costing or would cost in the future. For publishers and producers, the cost of archiving was still tied up in the cost of manufacture. This is also true of publications services where the archiving is considered an added benefit to the publishers who are served. Until several large archives have gone through at least one or two migrations or emulation developments, it will not be possible to separate the cost for the archives from the cost of doing business.

In the case where material such as electronic journals and Web sites are being archived that would otherwise have been acquired, it is possible to determine the cost. However, organizations such as OhioLINK consider this information to be proprietary. In this case, there is not only a cost for infrastructure, but a cost that compensates the publisher for the breadth and scope of the license agreements, as well as acknowledging that broader licenses mean fewer licenses and less overhead in paperwork and administration.

In cases where the library must play a more active roll in the archiving process, there is some anecdotal information beginning to be available. It should be noted that these efforts are relatively small to-date and, therefore, there has been limited consideration of the cost of new systems, disk storage, or of migrating across these technologies in the future. The major resource emphasis to-

date has been in the area of acquisition, collection development and intellectual property rights management. As part of its pilot project, the National Library of Canada estimated that it would have to reassign the equivalent of two full-time acquisitions personnel to handle an estimated workload of 500 to 1,000 new electronic publications per year. Since the number of electronic publication titles per year is expected to grow, the EPPP report called for the NLC to provide systems support that would help to streamline the process. At the National Library of Australia, there is no specific archiving budget for the PANDORA, but there are five staff members currently assigned to the project (Phillips, Personal communication. 1999.). At STIC (Taiwan), the limited operation is taking approximately 10% of a staff person's time (Flannery and Shuyu, Personal communication. 1999). This is estimated to increase to 50% when all journals in Taiwan are included in the archive.

For publishers who have gone to electronic production, it is difficult to separate the cost for the archiving from the cost of regular production. However, several of the publishers which serve learned societies, most notably AIP, are moving in the direction of providing archiving services for member societies that published their own journals. The cost model for this has not yet been determined, but it seems that this would need to be identified at some point in order to quantify the benefit of the service being provided. (Ingoldsby, Personal communication, 1999). AIP also provides physical copies of its archive to users at a minimal cost of $25-50 per issue. This includes the abstract and all supplementary information.

AIP noted that "the challenges are seen as coming from the inexorable growth of the journal-publishing endeavor. Costs in terms of both funds and staff time will rise year by year, and the annual charge will change from representing a small fraction of the total publishing cost to rivaling the costs of publishing the current volumes. Although AIP has been fairly successful in recent years in minimizing journal growth (to control subscription prices), there is a feeling among many publishers that online journals can be allowed to grow almost without limit. Archiving costs will be affected accordingly." (Scott, Personal communication, 1999)

The data centers reviewed that were most viable had operating budgets of $1M and averaged well over $2M. However, these are in narrow areas of science but with large disk storage demands. The director of the ORNL DAAC estimates that depending on the quantity of data and the infrastructure in place, it would require approximately $2M in start-up costs for an archive, with ongoing costs of approximately $2.5M. The LTER sites do not have specific information management budgets, but the Network director indicated that approximately 15% of each sites research grant funds (varying from) goes to information management. The proportion is higher for the Network Office. (Porter, Personal communication, 1999).

Comparing the cost of archives across data types has many pitfalls. However, it is significant to notice the dramatic increase in the cost with the increased complexity of multimedia. The start-up costs for DoD's pilot project is $23M over three years. To date the project has cost $22M to develop detailed specifications and install an unclassified video archival suite. The ongoing costs are projected to be $300,000 per year. (Borkowski, Personal communication. 1999.)

In addition to questions of start-up and ongoing operation, there is a serious issue of the long term financial commitment to archives. According to one data center director increasing recognition by scientific authors and funding sources is key to the success and sustainability of an archive. If a $2M data center goes forever it would need a $20M endowment or continuing appropriations commitment. The possibility of an endowment model was also raised by Clifford Lynch of CNI at a recent NSF workshop on this topic. (Lynch, Personal communication, 1999.)

Conclusions

Based on the analysis of the organizational models, the changing roles of traditional stakeholders, and best practices in digital life cycle management, general observations are made in the areas of most interest to ICSTI. These include policies, organizational models, and economic models.

Policies

To greater or lesser degrees, all stakeholders in the archiving and preservation chain are concerned about intellectual property. For many of the data centers, the issue is put in public versus commercial use terms, and is reflected in the types of access and services provided and the charges placed on them. For publishers and producers, intellectual property concerns are reflected in the kinds of business arrangements used to promote their archives. Intellectual property concerns have led some organizations to consider institutional archives, where the information remains under their control. Others, lacking the resources to do this, but still concerned about their intellectual assets, are contracting with publication services or trusted third-party repositories. Part of these contracts requires security and authentication on the part of the archive, as well as specific procedures for granting and continuing access. Libraries, consortia and users are increasingly attuned to intellectual property issues and their concerns for fair use in a digital environment are often reflected in the license agreements that are signed.

An area closely related to policies surrounding intellectual property is that of legal deposit of electronic publications. While lacking the legislation, many national libraries consider the collection and preservation of these publications to be part of their mandate. While continuing to raise awareness of the need for resolution of the legal deposit issues, national libraries in both operational and pilot projects (most notably Canada, Finland, Sweden and Australia) are getting rights holders permissions to collect and preserve these electronic publications. However, these issues are not resolved, and despite guidelines there are still many questions about what a particular national library should collect from an international publishing environment such as the Internet. In recently published guidelines, NLC "recognizes that there are inherent problems in applying a Canadian law like legal deposit in an international communications medium that does not necessarily recognize jurisdictional borders. For example, what would define a publisher as "Canadian" or a network site as "Canadian" given the ease with which networked sites can be mirrored and networked publications can be copied (i.e., many resources at a Canadian site may not be Canadian in origin) and given the volatility of network addresses (i.e., a document hosted at a Canadian site can be easily transferred to non-Canadian sites which are outside the

jurisdiction of Canadian laws)." (www.nlc-bnc.ca/pubs/irm/eneppg.htm) The same could be said for any national library depository.

Organizational Models

There are a large number of DEA projects, at various stages of implementation and operation. The most robust of these are in the area of numeric data, using the data center model. These centers are moving into increasingly distributed systems, where there can be increased heterogeneity in system and data architectures. The implementations in the nonnumeric data areas are not as mature. These archives tend to be more scattered with a variety of stakeholders doing prototypes and implementations of systems.

The simplest archives to achieve are those created by the publishers or large data producers themselves. The reason is that the publishers have more control over the formats and standards used. The producer and publisher can reduce the variety of incoming formats which eliminates much of the confusion related to multiple migration paths. Even if emulation is selected instead of migration, the fewer the number of software and hardware environments that must be emulated the easier.

However, from the standpoint of access and use, archiving by a single publisher does not provide the user with the "view of the world" that is needed or wanted. The scientific and technical information needed by the user may not reside in that publisher's or producer's archive. The subject may be too broad or too narrow. The user may be interested in a topic that is interdisciplinary and spans across multiple publisher niches.

There are several ways to provide more consolidation, at least in the text publishing area. Third party repository agents may continue or expand their efforts across disciplines. However, it is unlikely that they will be able to physically archive everything that could potentially be of value. Concerns are often raised about the cost of access, particularly for individual users, the overhead of such organizations, and the commercial nature of these efforts. If the information is not commercially viable, will these third-party organizations continue to retain it?

A single consolidated, global repository does not answer many of the concerns about the administration and understanding of the archive's contents. It is doubtful that one archive will fit all, at least in the near-term. Archives are likely to be organized around subject disciplines, since experts in the discipline are most likely to know the value of what must be selected, how it should be cataloged and how to make it accessible to various user communities. However, there is a sense in which the format or data type also comes into play. One cannot assume that an archive of numeric data can handle highly structured text equally well. In the short term, archives will continue to specialize both by discipline and by data type.

However, this will likely change as more scientific and technical information is made available in multimedia formats. Over the long term, more authoring tools, metadata creation tools, storage,

and access tools will be developed to better accommodate multimedia. The various parts of a complex document or of a multimedia work can be archived objects into smaller archives, that are built for the purpose of the particular data type. As the technology support for multimedia storage improves, there may be less emphasis on data type and more on the content and subject.

With a large number of models and increased interest in the future of digital information, many stakeholders are getting into some aspect of the archiving business. There are many organizations that appear to consider this a reasonable avenue for business growth, if not direct revenue generation, whether in support of electronic publishers or by providing the safety net that libraries, particularly consortia, are requiring for their electronic subscription investment. With the large infrastructure and varying skills needed to perform digital archiving satisfactorily, we may be seeing the rise of a new industry. Smaller publishers may continue to look for avenues by which they can contribute to one or more archives. However, for the third parties, publication production services, or even libraries that choose to archive digital material, it is often secondary to the provision of other services. The organization must be willing to "make the investment in infrastructure and have a purpose beyond archiving as a justification for doing this." (Sanville, Personal communication, 1999.)

In the meantime, is there a clear direction toward a standard model or set of models for DEA? What model is best? Is a standard model possible or desirable, given the fact that there is not a single archiving model in the paper world? With the exception of the numeric data centers, the archives are not large enough to provide evidence of what standards are emerging. The data centers are moving toward increased heterogeneity and ever looser federations. The other DEA organizations are already very heterogeneous in their approaches. Therefore, the organizational model for archives in the foreseeable future appears to be a loose network of archives covering special disciplines, geographic areas, or object types. Using network technologies and interoperable standards, the future model will likely be a network of disparate but interoperable archives. Individual communities are likely to develop standards and common practices; it is unlikely in this distributed environment that the same standards and common practices will suffice for all. Interoperability in a heterogeneous environment is likely to be the requirement.

The Open Archive Information System (OAIS) reference model, described earlier, appears poised to promote this interoperability beyond the realm of data centric archives. An OAIS that is robust enough to support the commonalities among the disciplines, data types and user communities, yet flexible enough to support the differences, is the most likely model to succeed.

Economic Models

Similarly, it is likely that there will be a variety of economic models for digital archiving. The economics vary from discipline to discipline. Some disciplines are more likely to reuse archival information for a longer period of time than other disciplines. This will impact not only the way the archives are managed and who archives them, but the value (and the cost) involved in retaining older materials. Some archives will be commercially viable, but others will not. Some

will need to charge for services, while others will not. When archives are governmentally appropriated, there is increasing recognition of a long term maintenance commitment, but there may not be sufficient definitive action and funding to support this recognition.

In addition to funding, the infrastructure, including coordination needed to make this work on a global scale, is the key issue to be addressed. A UK Public Records Office project has recently begun investigating the model of a coordinated network of archives. (A similar approach is being discussed by the Conference of European National Libraries (CENL) through a project called the Networked European Depository Libraries (NEDLIB). Funded by the EU with project leadership at the National Library of the Netherlands, this project is in the early stages of defining how these libraries would interoperate, retaining their autonomy and primary emphasis on the publications of their individual countries, yet providing improved access to all who need it. The NEDLIB project is significant in that commercial publishers, Elsevier, Springer and Kluwer, are acting as sponsors for the project.

Multiple Models in a Networked Environment

As the report of the ICSTI Electronic Publications Archive Working Group suggested, a hierarchy of archiving models with variant organizational and economic structure may be the initial DEA model (ICSTI, 1998). Even in the current projects, it is possible to see multiple, integrated models at work, that recognize a life cycle within the archiving function itself. For example, the central DOE site provides a catalog and coordinating function. Many of the DOE laboratories have their own archives which are also made available via the Internet. The OSTI site provides the "glue" for this network. The central site also provides backup for the distributed archives, when they decide it is no longer feasible to retain the item on the local archive. Similarly, the U.S. National Archives and Records Administration (NARA) and other U.S. government access activities such as the National Technical Information Service FedWorld system and the U.S. Government Printing Office Access system, provide access to the electronic archive of DOE through the InformationBridge product. NARA stands as the final archive depository, should the DOE OSTI program be discontinued. In a less formalized fashion, but equally distributed, several publishers, including AAAS, are contributing their content to multiple archives, with variant models, in addition to archiving their own electronic journals.

It appears the discipline specific, national and global archives will be built incrementally on the basis of pilot projects that lead the way and evolve into a complex network of content infrastructure. The issue has been recognized and the bandwagon is growing. In summarizing best practice areas, we see building blocks for future developments. The trick will be the coordination of these archives to reduce the expense of unnecessary redundancy and to tie the system together in an integrated fashion for the user.

Recommended Next Steps

Based on the survey and analysis conducted during this project, the following actions are

recommended for consideration.

ICSTI

1. **Many models are evolving and taking hold. Each stakeholder will be affected and the activities should be monitored for more specific and ongoing relevance to ICSTI member groups:**

- **Hold discussions on impacts of the various models (both organizational and economic) for classes of ICSTI members. Monitor projects selected by members to be models for their part of the industry, and provide opportunities for interaction between these projects and appropriate communities within ICSTI.**

Projects that include the specific stakeholder group or the portion of the information life cycle function in which a particular organization is interested should be monitored with specific reports back to ICSTI members interested in these particular areas. In addition to project monitoring, opportunities should be provided for interaction between the project managers of the selected projects and ICSTI members. The next annual meeting, or a special meeting cosponsored with ICSU, UNESCO or some other organization, would provide a forum for the discussion of these specific projects. It might also be valuable to hold the session concurrent with a major meeting where these projects might already be represented.

- **Interpret the draft Open Archive Information System (OAIS) Reference Model for the ICSTI Communities**

Since heterogeneity and a complex network seem to be evolving, the OAIS Reference Model is one worth further group exploration. It stands as a possible framework for data interchange needed across the various functions of an archive (regardless of the players involved), and across archives. However, the current reference model is still very data-centered. ICSTI should convene a small group or groups of stakeholders to interpret the reference model for the different communities -- primary publishers, secondary publishers, and libraries. During this process it should be possible to determine if the reference model has utility for a variety of stakeholders and a variety of data types. The CEDARS project in the UK has expressed an interest in working together with ICSTI on this review. This follow-on project should be done in the context of the ISO review of the draft reference model and should consider interoperability, standards, common practices and economic models that will have to coexist. The benefit to ISO and the Consultative Committee on Space Data Systems is that they will obtain a review by an expert community, outside the data community. The benefit to ICSTI is that it may find a model that can be used across its members and to inform the community at large.

- **Develop a Digital Electronic Archive Registry Emphasizing Digital Publications**

The Electronic Archive Registry, recommended by the ICSTI Electronic Publications Archive

Working Group, may act as a transitional mechanism between the current distributed, unintegrated archiving projects for electronic publications and the fully networked environment envisioned by the OAIS. The Working Group envisioned this registry as a finding aid for the location of where, by whom, in what format, and what parts of a publication are electronically archived. The data elements required for such a registry and the procedures whereby the registry is created, maintained and accessed must be developed. The Working Group suggested that the registry could be added to the ISSN system. The concept should also consider the work of other groups such as the Digital Object Identifier (DOI) Foundation and the national libraries/bibliographies.

- **Monitor and report on the key projects related to the cost and organizational issues of digital archiving**

This review has identified that there are still significant unanswered cost and economic questions related to long-term digital archiving. Some of these questions are related to the speed of technological change, while others are institutional. However, there are several significant projects under way that have been briefly identified in this report. They should continue to be monitored and progress on them reported to the ICSTI community. Recommendations for projects to be monitored include NEDLIB, the objective of which is the networking of depository libraries and the development of digital depository format standards for publishers; CEDARS, which is looking at the networking of UK archives; and Cornell University's Digital Library 2-Initiative which will address cost and organizational issues. Relationships should be established with these projects in order to learn about their progress and be able to report on the outcomes to the ICSTI listserv.

2. As appropriate, work at individual organization levels to promote digital archiving practices:

- **Recommend to ICSTI organizations that digital standards for metadata and object identification that are under consideration be reviewed with a particular eye to their ability to support long term preservation and access.**

In particular, work to ensure that the concept of archives and preservation is developed and used within existing and forming standards for metadata and identifier.

- **Provide testbed material for projects when possible.**

A significant way for ICSTI members to become involved and to learn more about the challenges and best practices in this area is to provide material for digital archiving testbeds. This is already being done by Elsevier, Kluwer and Springer in the NEDLIB project. There may be similar opportunities with other projects, including CEDARS and the Cornell University DLI-2 projects.

- **Promote multilateral projects, to promote the development of best industry practices**

in digital archiving

Promote round-table sessions at a follow-on ICSTI meeting that would bring together ICSTI members working on similar issues related to digital archiving so that resources, lessons learned, and pilot projects could be shared. Of particular importance would be discussions and pilot projects related to business models for digital archiving and intellectual property issues (particularly between national libraries and publishers).

Both ICSTI and CENDI

1. Make ICSTI/CENDI's interest in this area known so the organizations stay involved with the forefront of activities and continue to keep the debate visible with customers, suppliers, and funding sources.

- **Present a paper at the World Science Conference**

As suggested by the ICSTI Executive Board and planned in the proposal, the results of this study will be presented by Dr. David Russon at the World Science Conference in July 1999.

- **Develop a Statement of Concern regarding digital electronic archiving**

As many survey participants mentioned, the current projects in digital electronic archiving are often being done without adequate commitment and funding. There is concern that funding will not be sustained, and is not consistent with mandates to collect and preserve electronic information. As suggested by the ICSTI Working Group, ICSTI and CENDI should produce a Statement of Concern, either jointly or consecutively, that raises the issues of electronic archiving and continued preservation and access to these archives with stakeholders, policy makers and funding sources. Many of the stakeholder groups are represented by members of ICSTI and CENDI, and therefore, it should be in a unique position to "work through" this difficult task. As the ICSTI Digital Electronic Working Group indicated in its report, the statement should not only identify the need for and benefits to be gained by electronic archiving and continuing access, but it should identify guidelines for what constitutes an electronic archive and sufficient access. It should emphasize the need to support verbal commitments to digital archiving with proper programming and funding. The Statement of Concern should also identify further activities in which ICSTI and others can participate to ensure that the statement is put into action.

- **Publish an article on the results of the ICSTI/CENDI study**

While the report to the World Science Conference will provide some level of visibility for the efforts of ICSTI and CENDI as well as for the next steps necessary to move digital archiving forward, this will not reach all stakeholder audiences. It is suggested that an article be prepared from the study and published in a relevant journal. The investigators have already been approached by the editor of the *Journal of Electronic Publishing* for such an article.

- **Develop a topical area on either the open ICSTI or CENDI Web site that highlights digital electronic archiving.** (This could also be done as a joint effort.)

The topic of archiving was highlighted in the report from the June 1997 meeting and in a subsequent issue of the *ICSTI Forum*. Those documents, a summary of this report and other possible information gleaned from ICSTI members should be included as a special theme on the Web site. (There are many good sites that already address this issue, and there is no need to replicate them. However, links from a specific ICSTI or CENDI page to these other sites may be of value to ICSTI and CENDI members and others interested in this subject.) CENDI could consider highlighting this area as a special adjunct to the broader STI Manager part of its Web site.

This survey has emphasized that DEA issues require collaboration and coordination among a variety of stakeholders. There are numerous projects underway at many levels. The ICSTI and CENDI members can benefit from staying informed of ongoing activities. They also have experience and practical needs that can help to inform and move the state of DEA implementation forward.

> **Though we need to act, the compressed cycle between the manufacture of digital artifacts and the almost immediate imperative to preserve these same artifacts should not lead preservation decisions based on expedience. ...we should at least consider how future generations will come to place a value on a particular piece or collection. And by generations I do not mean a couple of generations measured by the computer/video industry, but for generations measured in human terms. - Paul Messier, Conservator, Boston Museum of Art**

Bibliography

American Association of Law Libraries, et al. "Principles for Licensing Electronic Resources: Final Draft. July 15, 1997. (www.arl.org/scomm/licensing/principles.html)

American Institute of Physics. "AIP Archive & Use Policy." June 12, 1998. (www.aip.org/journals/archive/arch&use.html)

"Australian Archives: Managing Electronic Records." (www.aa.gov.au/AA_WWW/AA_Issues/ManagingER.html)

Bantin, Philip and Gerald Bernbom. "Indiana University Electronic Records Project 1995-1997: Final Report to the National Historical Publications and Records Commission (NHPRC), 1998. (www.indiana.edu/~libarche/nhprcfinalreport.html)

Barkstrom, Bruce R. "Digital Archive Issues from the Perspective of an Earth Science Data Producer" (techreports.larc.nasa.gov/trs/papers/NASA-98-dadw-brb/Archival%201Issues.html#5.2 Scientific Data Archival Issues)

Barnes, John. "Digital Archiving: The Transition to Electronic Archives." Presented at Internet Librarian '98, November 5, 1998. (www.oclc.org/oclc/menu/ref_presentation.htm)

Beagrie, Neil and Daniel Greenstein. "A Strategic Policy Framework for Creating and Preserving Digital Collections." July 14, 1998. (ahds.ac.uk/manage/framework.htm)

Boyce, Peter. "Costs, Archiving, and the Publishing Process in Electronic STM Journals." Against the Grain, v. 9 #5,p. 86, Nov 1997. (www.aas.org/~pboyce/epubs/atg98a-2.html)

Brophy, Peter, Shelagh Fisher, Geoffrey Hare and David Kay. "Towards a National Agency for Resource Discovery Scoping Study." British Library Research and Innovation Report 58 [RIC/G/364], 1997. (www.ukoln.ac.uk/services/papers/bl/blri058/)

Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS): Recommendation Concerning Space Data Systems Standards." White Book CCSDS 650.0-W-4.0, September 17, 1998. (ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html)

Library Programs Service, U.S. Government Printing Office. "Managing the FDLP Electronic Collection: A Policy and Planning Document." October 1, 1998. (www.access.gpo.gov/su_docs/dpos/ecplan.html)

Garrett, John and Donald Waters. "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information." Commissioned by the Commission on Preservation and Access

and the Research Libraries Group, Inc." 1996.(www.rlg.org/ArchTF/tfadi.index.htm)

Getty Conservation Institute, the Getty Information Institute, and Long Now Foundation. "Time and Bits." (www.gii.getty.edu/timeandbits/index.html)

Haynes, David and David Streatfield. "A National Co-ordinating Body for Digital Archiving?" *Ariadne*, 15. May, 1998. (www.ariadne.ac.uk/issue15/digital/)

Haynes, David, David Streatfield, Tanya Jowett and Monica Blake. "Responsibility for Digital Archiving and Long Term Access to Digital Data." JISC/NPO Studies on Preservation of Electronic Materials. 1997.
(www.ukoln.ac.uk/services/papers/bl/jisc-npo67/digital-preservation.html)

Hedstrom, Margaret and Sheon Montgomery. "Digital Preservation Needs and Requirements in RLG Member Institutions." A study commissioned by the Research Libraries Group. December 1998. (www.rlg.org/preserv/digpres.html)

Helsinki University Library and Center for Scientific Computing in Finland. *Functional and Technical Requirements for Capturing On-line Documents (EVA-Project)*, No Date.

Heminger, Alan R. and Steven B. Robertson. "Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents" Nov. 21, 1998.
(tuvok.au.af.mil/au/database/research/ay1996/afit_la/rober_sb.htm)

ICSTI. "The Electronic Publications Archive - Report of a Working Group of the International Council for Scientific and Technical Information." December 1998..

ingenta, Ltd. "OCLC and ingenta in '2 for 1' Partnership to Offer Online Journal Services to Societies and Publishers." Press release, December 1998.
(www.ingenta.com/Tfedocs/press/oclc.html)

International Coalition of Library Consortia (ICOLC). "Statement of Current Perspective and Preferred Practices for the Selection and Purchase of Electronic Information." March, 1998.
(www.library.yale.edu/consortia/statement.html)

Joint Information Systems Committee & Publishers Association 'Model Licence' Between UK Universities and Publishers. January 21 1999.
(www.ukoln.ac.uk/services/elib/papers/pa/licence/Pajisc21.html)

Kahle, Brewster. "Preserving the Internet." *Scientific American*, March, 1997.
(www.sciam.com/0397issue/0397kahle.html)

Kelly, Maureen C. "The Role of A&I Services in Facilitating Access to the E-Archive of

Science." *ICSTI Forum* No. 26, November, 1997. (http://www.icsti.org/icsti/forum/fo9711.html)

Kuny, Terry. "The Digital Dark Ages? Challenges in the Preservation of Electronic Information." International Preservation News, No. 17, May 1998. (ifla.inist.fr/V1/4/news/17-98.htm#2)

Messier, Paul. "Observations on UPF as a Conservation Medium". Paper delivered at the Association of Moving Image Archivists Convention, December 9, 1998 (info.wgbh.org/upf/papers/Messier.html)

Meyers, Barbara and Linda Beebe. "Archiving from the Publisher's Point of View." A white paper prepared for Sheridan Press. 1997.

NASA DAACs. "Data and Information Services for Global Change Research." (ivanova.gsfc.nasa.gov/daac/fliers/m_daac.html)

National Library of Australia. "Selection of Online Australian Publications Intended for Preservation by the National Library of Australia." (www.nla.gov/au/scoap/guidelines.html)

National Library of Australia. "Preserving Access to Digital Information (PADI)." (www.nla.gov.au/padi/)

National Library of Canada, Electronic Collections Coordinating Group. *Networked Electronic Publications Policy And Guidelines*, October 1998. (www.nlc-bnc.ca/pubs/irm/eneppg.htm)

National Research Council. *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources*. 1995.

Natural Environment Research Council. "Environmental Data - a Key Resource" (www.nerc.ac.uk/environmental-data/)

Okerson, Ann. "The Current Licensing Landscape: Does It Scale?" Presentation at the National Federation of Abstracting and Information Services Annual Meeting, Feb. 22, 1999.

Oxford University. "Preserving the Electronic Assets of a University." (users.ox.ac.uk/~alex.hfs.AXIS-paper.html)

Pack, Thomas and Jeff Pemberton. "A Harbinger of Change: The Cutting Edge Library at the Los Alamos National Laboratory." *ONLINE Magazine*, March 1999. (www.onlineinc.com/articles/onlinemag/pack993.html).

Rothenberg, Jeffrey. "Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation."January, 1999. Report to CLIR. (www.clir.org/pubs/reports/rothenberg/contents.html)

Sharpe, Louis, D. Michael Ott, and Carl Fleischhauer. “Library of Congress Manuscript Digitization Demonstration Project: Final Report.” October, 1998. (memory.loc.gov/ammem/pictel/index.html)

Shepard, Thom. “Universal Preservation Format Update.” D-Lib Magazine, Nov. 1997. (http:/www.dlib.org/dlib/november97/11contents.html)

Shepard, Thom. “Working Draft of the UPF Recommended Practice.” February 2, 1999. (info.wgbh.org/upf/)

Smith, Wendy. “PANDORA - Boxing for Survival: Archiving, Preservation and Access Issues Related to Australian Internet Based Publications.” Paper presented at the ‘On the Edge Conference’, Perth, October 1997. (www.nla.gov.au/nla/staffpaper/wsmith3.html)

Society of American Archivists. “Statement on the Preservation of Digitized Reproductions.” June, 1997. (www.archivists.org/governance/resolutions/digitize.html)

Tombs, Kenneth. “Report to the Data Archival and Information Preservation Workshop.” Washington, D.C. March 26-27, 1999.

Uhlir, Paul. “Framework for the Preservation and Permanent Public Access to USDA Digital Publications”, November 1997.

U.S. National Archives and Records Administration, Electronic Records Center. “Managing Electronic Records, National Archives and Records Administration Instructional Guide Series” (gopher://gopher.nara.gov/00/managers/federal/publicat/elecrecs)

Waters, Donald. “Toward a System of Digital Archives: Some Technological, Political and Economic Considerations: The Rearranging Effect ” Presented at the Association of Research Libraries Meeting , October, 1998. (arl.cni.org/arl/proceedings/131/waters.html)