APPENDIX 20. LINKING THE INFORMATION LIFE CYCLE CONCEPT WITH DIGITAL LIBRARIES

NCLIS POLICY RECOMMENDATIONS ON LINKING THE INFORMATION LIFE CYCLE CONCEPT WITH DIGITAL LIBRARIES

Written by Satadip Dutta, Virginia Institute of Technology
Reviewed by Edward A. Fox and Shalin Urs[16]

EXECUTIVE SUMMARY[17]

The aim of this paper is to provide recommendations that would be used in the study of reforms to the Federal Government's public information dissemination laws, policies, programs, and practices. The paper describes the information lifecycle that outlines the process of creation, retrieval, and utilization of information. Salient features of government information are then discussed. The paper then explores the issues relevant to government information that include distribution of government information artifacts, issues in standardization of publishing formats, remodeling the information publishing model, and digital preservation of these artifacts. The scope of the paper is limited to exploring issues and making recommendations related to government information artifacts that would help the scholarly community.

THE INFORMATION LIFECYCLE MODEL

Information stored in libraries passes through a definite lifecycle that involves major phases like:

- Information Creation: This phase primarily targets the creation of any kind of information. It involves authors and other creators actually preparing and modifying the information. The organizing and indexing of this information to facilitate retrieval in later phases is also a part of this phase.
- Information Search: This phase deals with retrieving the information stored in (digital) libraries and other repositories. Activities like distribution of information also may be involved in ensuring widespread retrieval of relevant information.
- Information Utilization: This phase deals with issues in accessing the distributed warehouses of information. Information selected may be utilized for creation of new information. Issues related to preservation and mining of information also are relevant to this phase.

Figure 1 illustrates the lifecycle of information. This portrayal was prepared in 1996 at an NSF-funded workshop on Social Aspects of Digital Libraries, hosted by the Department of Information Science, University of California, Los Angeles. The report proposed a definition of digital libraries that encompassed two complementary ideas:

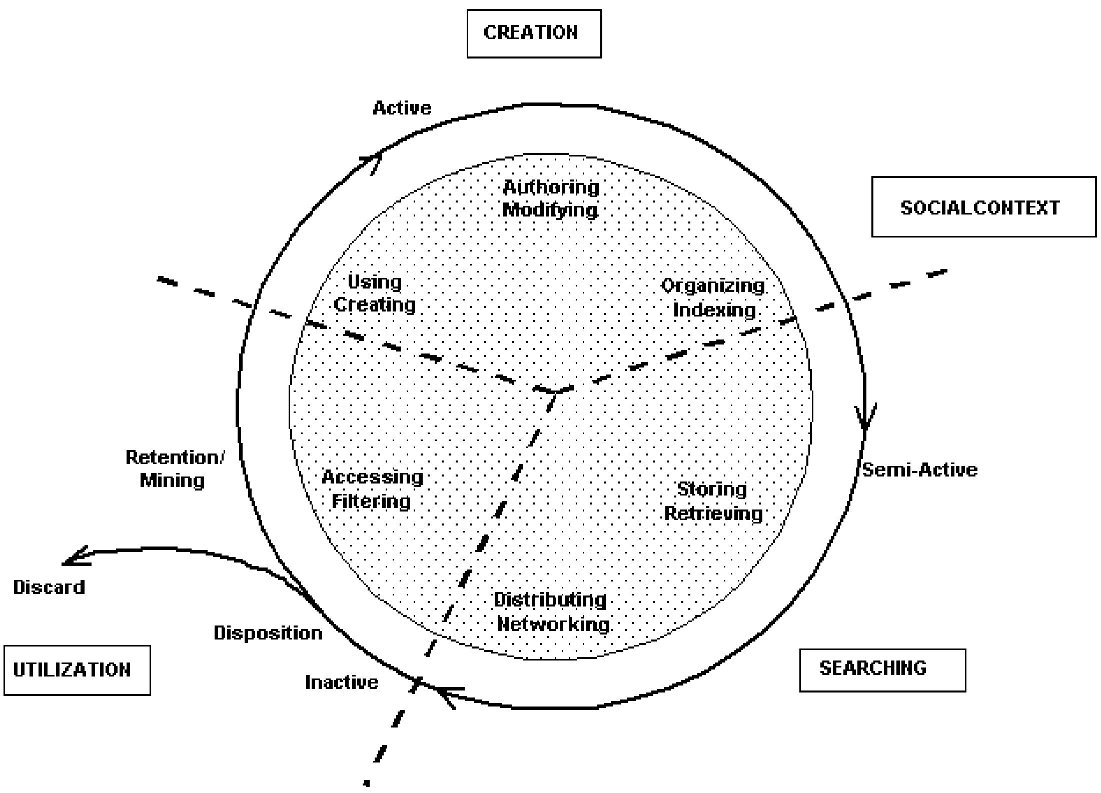
[16] Created by Satadip Dutta for Virginia Tech CS6604 term project ("Digital Libraries"); reviewed and edited by E. A. Fox (Professor) and Dr. Shalini Urs (visiting Fulbright scholar).

[17] Available at http://www.nclis.gov/govt/assess/assess.appen20.pdf. Interim revision October 16, 2000; final revision November 28, 2000.

1. that they extend and enhance existing information storage and retrieval systems, incorporating digital data and metadata in any form;
2. that digital library design, policy, and practice should reflect social context.

Creating, seeking, and using information are socially situated human activities. Some activities may evolve in the predicted directions as defined by the information lifecycle. However, there also are many less regular information activities that: switch back and forth between phases, skip phases, or end before the cycle is complete.

Information Life Cycle



NOTE: The outer ring indicates the life cycle stages (active, semi-active, and inactive) for a given type of information artifact (such as business records, artworks, documents, or scientific data). The stages are superimposed on six types of information uses or processes (shaded circle). The cycle has three major phases: information creation, searching, and utilization. The alignment of the cycle stages with the steps of information handling and process phases may vary according to the particular social or institutional context.

Figure 1: Information Lifecycle[18]

[18] Adopted from Christine Borgman, editor. *NSF Workshop Report on Social Aspects of Digital Libraries*, 1996, http://www-lis.gseis.ucla.edu/DL/.

Details relative to the information lifecycle with respect to particular types and collection of publications may vary according to the particular social or institutional contexts. Government information has distinct properties that are not always present in other information artifacts. For example a government information artifact is always produced in a context related to the political, technical, administrative, legal, and temporal setting. A context can be defined as a certain time-delimited environmental and social state. For example a foreign trade policy forged during a time of war should be understood in the context of war. The information artifact may not, however, contain information about that context. If contextual information (e.g., being at war) is not captured, the significance and rationale behind document creation may be lost when the document is viewed at other times (e.g., in times of peace). Therefore production and publishing of government information artifacts involves not only the development of the material but also recording the context of production (e.g., in metadata or a hyperlinked document).

INFORMATION CREATION

This section contrasts the current information publication model with a digital library based scheme. Creation of metadata to facilitate information retrieval, and some issues related to document standards, are then discussed.

Information Publishing Model

The current information publishing process goes through a series of phases. Beginning with author, it moves on to the editors, then to publishers, and thence to catalogers and librarians, who add value and enable published information to be consumed by the general public. Figure 2 illustrates this sequential model and describes the steps required therein for authors and readers to communicate over space and time.

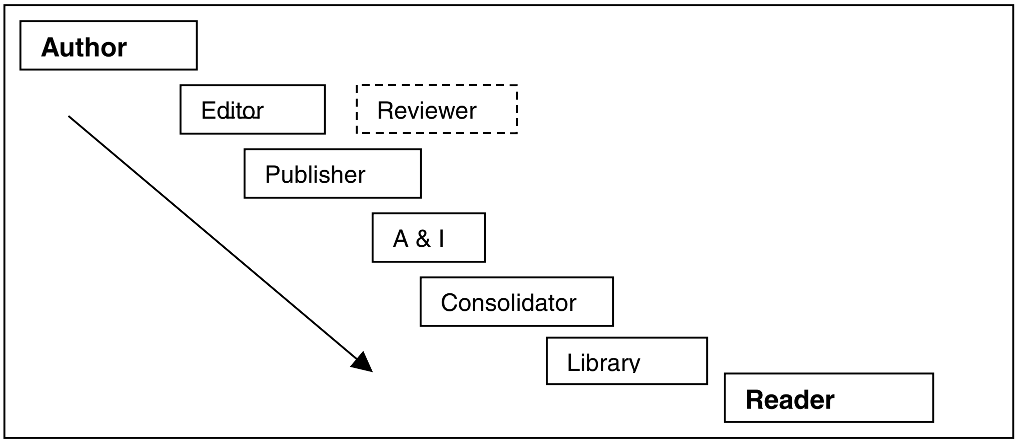


Figure 2: Sequential Information Publishing Model

A different model (see Figure 3) may result when all communication occurs in the same cyberspace (e.g., Internet), or, equivalently, in the same (federated/distributed) digital library. Participants in the process, regardless of their role, may simply be thought of as "users", who play different roles at different times. This "users direct" approach allows submissions to become available at point/time of creation, perhaps with improved/approved versions resulting later. The current revolution in electronic publishing also allows us to aggregate users/roles differently. Using Internet terminology we have:

authors/creators, data providers, and service providers. In this model the data provider is responsible for managing collections/archives that follow content creation.

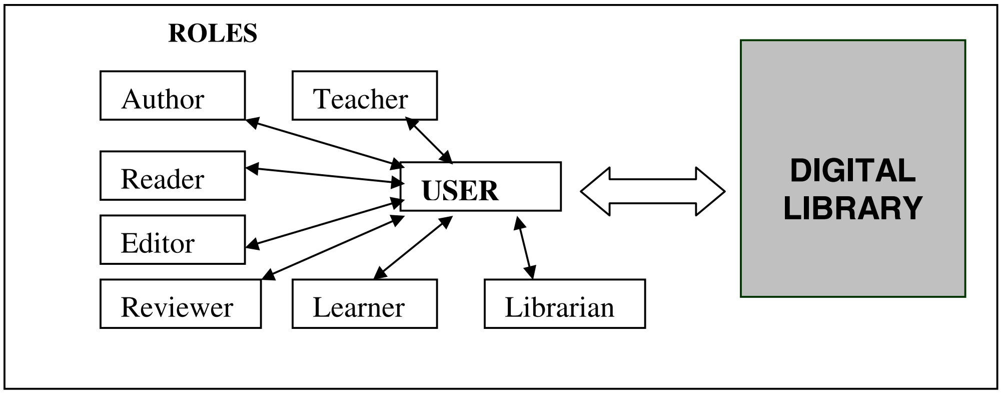


Figure 3: Internet Enabled Publishing Model

In the Open Archives model,[19] data providers need only support a simple harvesting protocol and provide extracts of metadata in a common minimal-level format in response to requests from service providers. The Open Archives Initiative is currently formulating an interoperability framework that would support both e-prints and a wide variety of other types of (scholarly) data archives. Service providers use extracted metadata to build higher level, user-oriented services, such as catalogs and portals to materials distributed across multiple content-bearing sites. Figure 4 illustrates service provision and data provision as the two main aspects of the digital library. Thus from a sequential model, we may shift to an Internet enabled publishing model where the participants, their roles, and the distribution of responsibilities may differ.

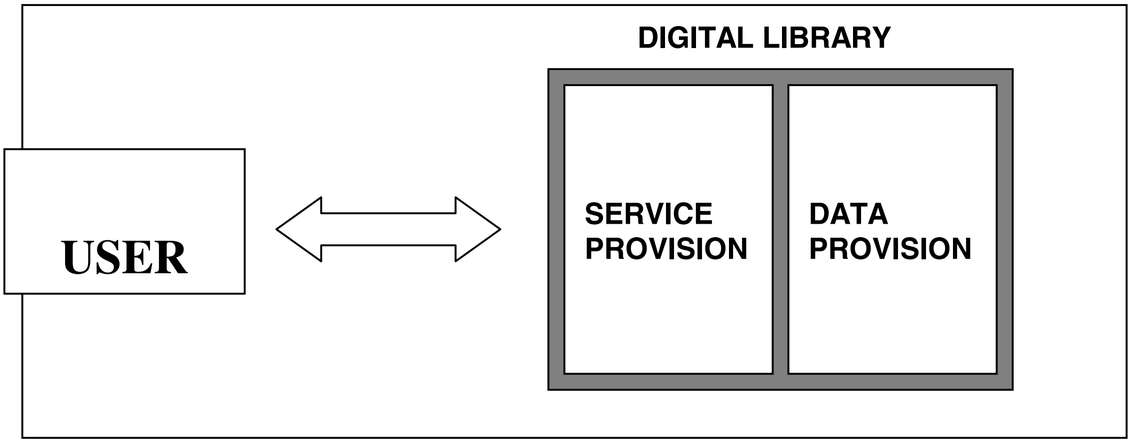


Figure 4: Service Provision Using Metadata

Figure 4 illustrates that users create, interact, and use data through a layer of services. These services may enable the user to create/modify information artifacts as described in Figure 3.

[19] The Open Archives Initiative, OAI, was launched Oct. 1999 and aims to support interoperability of archives. The current emphasis relates to a harvesting protocol and architecture of very simple data collections ("archives") that have digital objects, metadata objects, and support the protocol. It focuses on a middle, or harvesting, layer, assuming a lower document model layer, and services in layers above. Open Archives Initiative, web site, http://www.openarchives.org/.

Metadata Repositories

The move from a sequential information model to an Internet enabled publishing model may shift more of the responsibility for metadata and index creation to the creators of information artifacts and their agents. Metadata may be stored in repositories that can be used when searching for information present in libraries. The creation of metadata is extremely important for digital objects like datasets, music, speeches, video, maps, and pictures (though improving content-based multimedia information retrieval may play an important new role). The scholarly community also may exploit metadata repositories to mine information, e.g., for discovering trends. Government or commercial (public and private) firms may develop and apply different retrieval routines. These may extend search capabilities beyond keyword-based models to support retrieval by describing the semantics or context of the information. Superior indexing techniques will also play an important role. Research focusing on techniques to automatically index multimedia information, plus semantics and context, with minimum manual intervention, should be encouraged.

Document Formats

Changes in information publishing may necessitate interoperable formats and standardization. In the new model the author/creator may be responsible for submission of information artifacts in standard forms to the digital library. Today there are standards like HTML and PDF. Other formats will evolve and prove better. For example the emerging XML standard can be used to represent, interchange, and manipulate a wide variety of data and information artifacts. Digital libraries also may use such schemes when migrating existing documents, which now exist in a variety of formats. This problem primarily exists due to the lack of consensus about document publication formats.

Apart from the migration and conversion problem, the material produced for government by the research community also needs to be distributed effectively so that everyone can easily find desired information. Many of the document formats require special readers that people may or may not have access to easily. These tools may be available free of cost but the existing infrastructure may not be sufficient to grant access to everyone. Users should be able to view the files without any special requirements. Therefore the software (like readers) must be made publicly available and steps should be taken to ensure that every computer has necessary software installed. This can help eliminate the problem of users with disparate backgrounds having different access capabilities.

INFORMATION SEARCH

This section discusses the creation of digital libraries on the scale of national libraries. The possibility of using services of non-government agencies to build effective retrieval mechanisms and the issues of registering information are then explored.

Digital Libraries

People have traditionally viewed libraries as repositories of information that are easily identifiable and accessible. The creation of very large (e.g., coordinated national) digital libraries necessitates the need to provide the same or greater volume of information, along with ease of access. To make the digital libraries well known or truly identifiable it is necessary first to at least promote them as alternatives to and extensions of the various conventional libraries. This may involve physical creation of multiple locations of digital libraries that are interconnected. Distributed or federated digital libraries are now popular for this and other social/economic/political reasons. Creation of multiple sites with mirroring/replication balances the load and results in better performance. Focused promotion activities can help publicize such digital libraries.

However, digital libraries will attract tremendous use, probably several orders of magnitude more than conventional libraries, simply because they can be superior to current systems and services. They can seamlessly handle all media types. They should seamlessly handle integration of wide varieties of data and information, with powerful and tailorable services. Not only should there be highly effective searching, but also browsing, linking, navigation, summarization, visualization, routing, filtering, and support for new types of artifact-supported collaboration and communication. They should improve with changes in the emerging networked world, yet provide continuity with the past, building on traditional values.

Thinking of a library as institution, there may be value in the concept of a US National Digital Library, supporting search and other services. NSF is developing NSDL, the National Science (Mathematics, Engineering, and Technology Education) Digital Library,[20] to open Fall 2002, and this has caught the imagination of educators and will have a profound impact on education in the nation. The California Digital Library may have some of the same effect in that state. If the scope of the Library of Congress and National Archives will stay roughly the same, it seems that there is incomplete coverage in the US in our current situation, though we do have Library of Congress, National Library of Medicine, National Agricultural Library, NSDL, etc. There is no digital library covering all fields (with respect to what is generated by the government, deposited according to its laws, or collected by it), even virtually.

Registration

In the United States, information about various topics is collected by different agencies. Often, the agencies that collect these data work independently of each other. This leads to difficulties regarding "registration". The registration problem arises when there is no way to align data for proper organization and integration. Ideally, different types of information can be aligned to produce multiple different views or perspectives that may not be evident from a single document. For example the percentages of different illness affecting children who are 5-9 years old may be collected by one government agency. Reports about the levels of various metals in the soil may be produced by another agency. There may be a possible link between the presence of various metals in the soil and the weakened immunity of children of a particular age group in that area. To illustrate further, a government agency may supply maps and other cartographic information about a particular locality. There might be another agency that produces information about the layout of utility lines, water pipes, drainage system, and the like for a given area. Before undertaking some repair work for the drainage system there may be information that could be derived from aerial photographs of the locality. This might lead to shorter decision times for servicing and repairing drainage systems. The absence of any form of registration makes the extraction of these types of conclusions almost impossible. Digital libraries may benefit from frameworks, using metadata, standards, and conventions, which allow registration of information artifacts.

Retrieval Mechanisms

Once the information is stored in a distributed manner across the country and once registration issues are resolved, the next step would be to create mechanisms that retrieve relevant information for users. These mechanisms require the creation of suitable user interfaces for all segments of the population. For example a scientist searching for recent speeches by Nobel Prize winners in physics might also

[20] The NSDL site is http://www.smete.org/.

wish to look at related publications. In other scenarios users may try to find information on topics for which they have little background. Further, interfaces should not only try to present the information retrieved but also provide mechanisms that allow users to restrict the context of the information artifact. Otherwise, if a person searches for stars, the digital library might retrieve documents related to entertainment stars, astronomy, songs that contains stars in their lyrics (like the Star Spangled Banner)—all very different contexts.

Retrieval mechanisms also need to be supported for the next generation of mobile computing devices since they increase the accessibility of information relative to the various facets of daily life. Such approaches, along with improved kiosks and programs to improve access in schools and public libraries, may complement ongoing efforts to eliminate the digital divide.

INFORMATION UTILIZATION

This section looks at the issues related to preservation of information artifacts. This is extremely important in the information lifecycle because information artifacts generated in the future need to refer to information produced in the past.

Preservation

Documents need to be preserved such that the context and the history behind the creation of the document are stored in ways that make it easy to retrieve and comprehend. This would allow people in the future to effectively evaluate reports that were produced in the past, with the correct perspective. Also, digital libraries have to cope with the migration of the documents stored in older formats to newer formats.

There continues to be a rapid change in technology that makes digital media obsolete very quickly. For example, as new storage formats evolve, storage capacity increases but at the same time the playback devices for older media become obsolete.
Digital storage media are usually fragile compared to paper. Therefore to maintain a collection in digital media regular checks of the information artifacts become necessary. Also it becomes important that the data is mirrored at certain remote locations. This gives rise to legal and copyright issues, as the information artifacts need to be periodically copied. For example, the Internet Archive (www.internetarchive.org) today plays a valuable role in archiving the Internet. But it may not actually follow the letter of the law since it is a business drawing upon copyright materials, without authorization from copyright holders.

Some extremely pertinent points are raised by the report on *Digital Strategy for the Library of Congress*[21] in this regard. The preservation responsibilities can be classified into loosely defined categories like:

1. a creator, active collector, and primary custodian for digital information artifacts;
2. a partner in preserving distributed digital collections.

[21] National Research Council, Computer Science and Telecommunications Board. *LC21: A Digital Strategy for the Library of Congress.* Washington, DC: National Academy Press (2001). This reference is to a prepublication copy, dated July 26, 2000. http://www4.nationalacademies.org/news.nsf/0a254cd9b53e0bc585256777004e74d3/bd6c8fce95b00a6d852569280047753a?OpenDocument.

Initiatives in the direction of assigning stable, long-term responsibilities to organizations like the Library of Congress would help in preservation activities for digital libraries.

CONCLUSION

This paper looks at the information lifecycle in the context of modern technology. It discusses aspects of Internet enabled publishing and other changes facilitated by digital libraries. It recommends further support of research to improve retrieval, as well as of mechanisms to reduce the digital divide. Further, it highlights key requirements, e.g., that government information artifacts must have their content and context preserved.