The University of Chicago Library Digital Archiving Program:

Scope, Mission, and Functions

with

Recommendations for Phase One

The University of Chicago Library
February 2005

Table of Contents

Establishment of a Digital Archiving Program

Context for Digital Archiving

Materials in digital format are becoming an increasingly significant part of our intellectual and cultural heritage. The fugitive and complex nature of these digital materials and the challenges of maintaining their integrity and accessibility over time are of continuing concern. Hardware and software changes leave files unreadable in a relatively short time. Additionally, digital technology enables library, archive, and museum collections to include materials made accessible via the Internet. Long-term preservation of these materials has been recognized as a national and international concern. For example, the Library of Congress has been charged to develop a National Digital Information Infrastructure for Preservation (NDIIP), the National Archives has undertaken creation of an Electronic Records Archive (ERA), and the Consultative Committee for Space Data Systems (CCSDS) has developed a reference model for an Open Archival Information System (OAIS).

The library profession is taking a leadership role in addressing the many challenges that are a part of digital archiving. The Digital Library Federation provides a context within which to collaborate on the development of standards, tools, policies and best practices. As a custodian of intellectual and cultural heritage, the University of Chicago Library has a significant stake in preserving a wide variety of digital materials on a long-term basis. These materials include: journals, books, images, datasets, and other commercially-published information; potentially ephemeral electronically-produced material such as organizational reports or subject-based websites; government documents; products of University of Chicago faculty including research notes and datasets, web-based projects, class materials, drafts, publications, and correspondence; University of Chicago publications and administrative records; email and attachments; and an increasing number of University of Chicago Library-produced web sites, databases, and other material. The Library's role is not only to make digital materials accessible now, but also to fulfill its traditional responsibility as a long-term repository for research resources.

Program Recommendation

The University of Chicago Library already has a well-established infrastructure for carrying out preservation and archiving activities for its print-based materials; however, it has not developed a comparable infrastructure for digital materials, which require intervention at an earlier stage of their life-cycle, attention to the use of non-proprietary formats and standards, and research into emerging tools and best practices. A management and technical infrastructure is needed that involves budget lines, hardware and software, organizational responsibilities, and establishment of policies, procedures, and best practices in a dynamic environment in which some standards are still emerging. Commitment to a new program of this scale will require significant new resources. A program to preserve and archive digital materials will involve the creation of new partnerships with Networking Systems and Information Technology (NSIT), the University of Chicago Press, and others on campus, as well as continued participation in national initiatives. The University of Chicago Library Digital Archiving Group recommends that this Library establish a program to ensure that digital information with lasting value is preserved for future access.

Program Scope

In defining a digital archiving program scope to fulfill this mission, many factors must be taken into account, including University administrative policy, organizational structure, and budgetary support. Within this University context, the definition of scope should focus on materials of interest to the Library because of their long-term intellectual, cultural, legal, operational, and community values.

Items selected for long-term archiving will be a subset of the materials for which there are also short-term intellectual, operational, and legal preservation needs. The Library will collaborate with others in the University on short-term management and maintenance activities in order to ensure long-term preservation goals can also be met. The Library will undertake direct responsibility for archiving some of the materials described below and for others will work collaboratively to address both short- and long-term archiving needs. Digital materials which are unique, and for which we are the primary steward, will be given high priority for direct management. Digital materials in which stewardship is shared, or which are not unique to our institution, will be addressed collaboratively both within the University and in conjunction with the national and international library community.

The following categories of material fall into the scope of the Library's digital archiving activities:

A. Digital materials that are locally created, exclusive or unique to the University, currently collected by the Library, and worthy of selection for long-term preservation

- Library-produced materials, including catalogs, bibliographies, databases, web sites, and digitized collections
- Materials created by faculty members, and other individuals and groups associated with the University, that constitute collections of papers and records
- Institutional and administrative records, including official publications, that are selected for inclusion in the University Archives

B. Digital materials that are locally created, exclusive or unique to the University, and not currently being collected by the Library, but which are worthy of selection for long-term preservation within a digital archive

- Records and data produced by academic and non-academic University management systems
- Web-based materials providing University services or describing University activities
- Electronic mail related to University activities
- Electronic instructional materials
- Materials constituting teaching and research collections that are created and managed at the University
- Data sets produced by faculty- or student-directed projects
- Materials that are published by the University of Chicago Press

C. Digital materials not exclusive or unique to the University, for which the Library provides access, and which the Library has a shared, inter-institutional interest in preserving

- Materials that are freely available on the internet
- Licensed materials that are accessible but not owned or locally stored by the Library
- Commercially produced materials that are locally stored by the Library

University of Chicago Library Digital Archiving Program

Program Mission Statement

The digital archiving program supports research, teaching, and administrative needs of the University of Chicago community through the preservation of digital materials for future access.

In pursuit of this mission, library staff members responsible for the program

- establish and maintain a sustainable infrastructure for submitting, storing, and disseminating digital materials according to accepted standards and best practices
- advise creators of digital materials on appropriate standards and best practices
- establish criteria for the acceptance of materials into the digital archive
- provide for discovery and access to archived digital materials in accordance with rights policies
- collaborate across the Library in defining and evaluating the scope of the digital archive
- collaborate with other members of the University community in addressing issues of digital archiving
- seek opportunities for collaboration with other institutions and national and international initiatives that advance the program
- inform and educate the Library and the University community about the digital archiving program
- monitor and assess commercial and non-commercial developments that could benefit the digital archiving program
- engage in research and development activities that advance the goals of the digital archiving program

Implementation of a Digital Archive

Functions of a Digital Archive

A successful digital archive provides the functions required to store and manage digital objects within the archive, as well as the functions that allow users to interact with the contents of the archive. Storage and management functions must guard against risks of loss due to hardware failure, software obsolescence, format obsolescence, and security breaches. Service functions must support user interaction with the archive within an identification, authorization, and authentication framework. These archival functions are effected by a combination of technical infrastructure, archival policies, and management. The development of classes of digital objects and classes of users will facilitate batch operations and automated processing.

I. Service Functions

Service functions allow for individuals to interact with the digital archive. The level of functionality a user will experience will depend on a combination of the class to which that user belongs and the rights associated with particular digital objects or classes of objects.

Deposit/Ingest: Deposit into the digital archive requires an authentication process and mechanisms for deposit of digital data with their associated metadata. Policies are needed determining what kinds of data and metadata may be deposited and by whom. The archive should support the deposit of simple, compound, or complex objects either individually or by batch process. The archive should record the date of deposit and any relation to other items in the archive *(e.g., if an object is an updated version of another object)*.

Discovery: Discovery functions allow users to identify what objects are stored in the digital archive, though not all users may be allowed to discover all objects at all times (either data or metadata). Objects may be restricted from discovery for reasons such as copyright status or privacy concerns. Restrictions on an object may change over time and will typically be defined with reference to date spans. The ability for a user to discover an object does not necessarily mean they can also access that object (*see dissemination*). Metadata records must maintain information about restrictions on a digital object including to whom the item is restricted and for how long. The digital archive should provide for discovery of individual items as well as various classes of items.

Dissemination/access: The archive's dissemination function must calculate accessibility depending on the combination of the class of the user and the restrictions placed on the object or class of objects. Objects may be restricted based on copyright status, university policy, depositor restrictions, or archive policy. Restrictions on an object may change over time and will typically be defined with reference to date spans. The archive must maintain information about restrictions including to whom

the item is restricted and for how long. The archive should disseminate objects with the metadata sufficient to allow a user to understand the provenance of the object and to successfully render the object.

Deletion/withdrawal: Items that have been deposited in the archive might, under certain circumstances, later be withdrawn/deleted. More typically, a replacement object is deposited and the older version is also retained. Policies are needed to govern when and by whom an item can be withdrawn. The archive should record the withdrawal of an object and the reasons for doing so.

II. Digital Object Storage and Management Functions

Storage and management functions may be performed upon deposit in the archive or at specified intervals. Frequency of such intervals may be set by policy or negotiated with a depositor. The digital archive should provide for automated storage and management actions based on policies *(e.g., a fixity check performed periodically)* and/or recorded metadata *(e.g., access to the resource done only after date specified in the object's metadata)*.

File Compression: Lossless compression can be done on archival files in order to conserve space. Policies may be established to determine which classes of files may or may not be compressed using which compression schemes. The archive should record any compression actions that have been applied to a file and any agreements made with a depositor as to when and if compression is acceptable.

File Migration: File migration is performed in order to allow files to continue to be useable as hardware and software change over time. Migration may not preserve all characteristics of the original document and so the essential characteristics of the digital object must be identified on deposit. Successful file migration relies on accurate identification of file formats in the archive *(see also validation and normalization)* and is simplified by policies describing the essential characteristics of various classes of files. The archive should record file format information, any migration processes that have been done to a file, and any information provided by the depositor regarding essential characteristics for a file or a class of files.

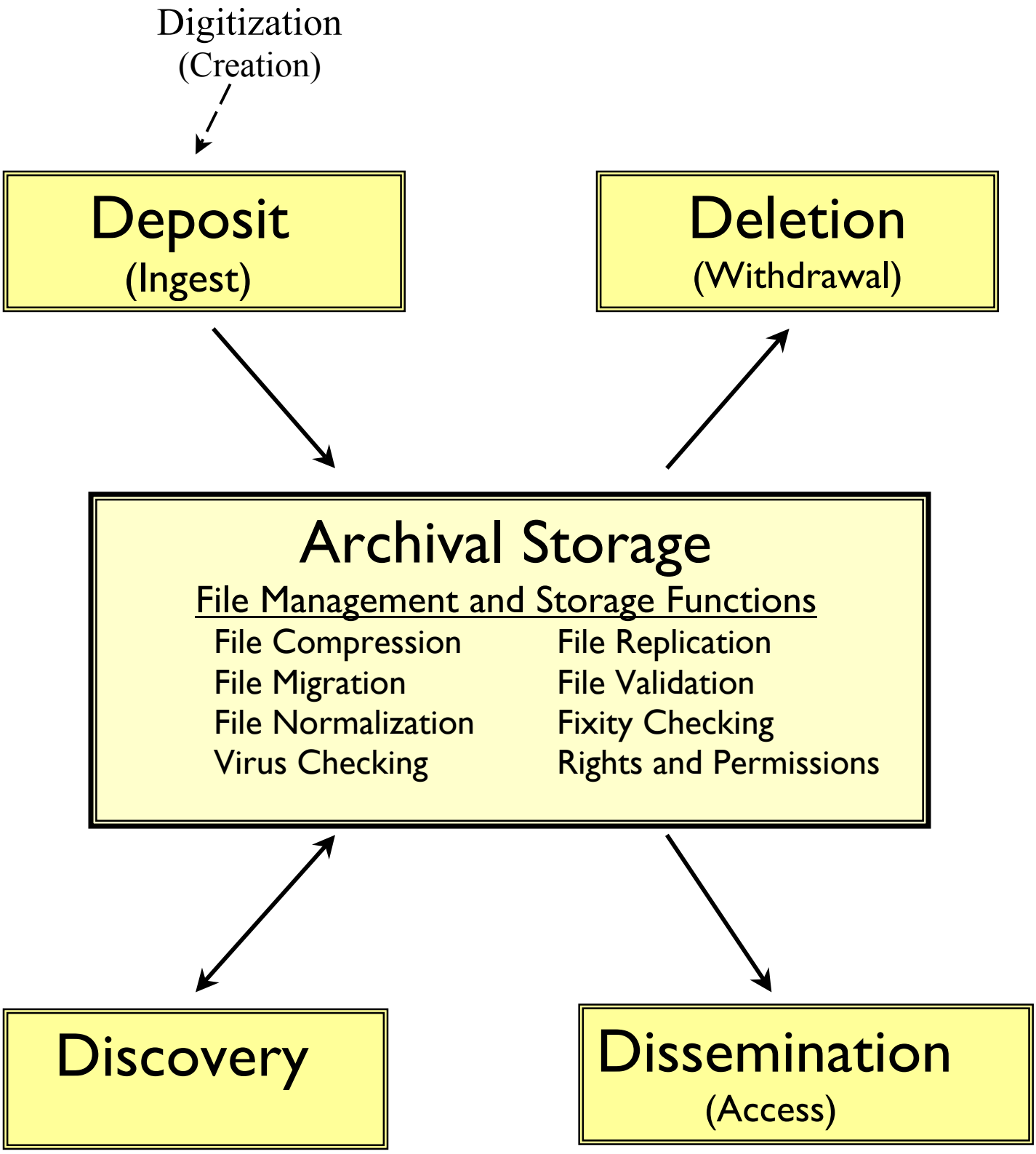
File Normalization: Normalization of files, when possible, ensures that files conform to documented, open standards so that data can be renderable in the future when a program or platform is no longer readily available. Policies are needed regarding recommended file formats and normalization processes, and whether files can/should be normalized upon deposit if not already in one of these recommended formats.

File Replication: File replication and storage of duplicate files in physically separate locations guards against file loss resulting from disk failure, file corruption, fire damage, or other localized disasters. Policies need to be established to ensure that replicated files stay within the control of the archive, to

ensure that unwanted dissemination does not occur in violation of an agreement with depositors or intellectual property rights.

Fixity, validation, and virus checks: These processes ensure the stability of the files in the archive by checking that a file is what it purports to be, and that it has not been corrupted over time. The archive should record when these functions have been performed on a file and any associated information that needs to be maintained in order to provide these functions over time.

Digital Archive



Recommendations for Phase One

The Implementation Group recommends a phased approach to the development of the Library's Digital Archiving Program. Phase 1 will build a system to register users, build interfaces for the deposit of selected data and metadata into the Archive, and define and develop interfaces for the discovery and dissemination of each type of document identified in the Proposed Content listed at the end of this document. Phase 1 will also define data management functions and withdrawal interfaces that will be built in Phase 2. Significant groundwork has already been laid by the Library's working groups and outside initiatives, such as PREMIS, which will inform the design and implementation of Phase 1. These components (described more fully below) will be built in Phase 1:

1. User registration system
2. Digital content inventory
3. Core metadata element set
4. Deposit interfaces
5. Dissemination interfaces

Phase 2 will test the deposit of greater numbers and types of documents into the Archive and build the data management functions and withdrawal interfaces that were defined in Phase 1. Phase 2 will also promote the use of the Archive, educate users and monitor costs of maintaining and upgrading the Archive. These components (described more fully below) will be defined in Phase 1 and build in Phase 2:

6. Archival actions for deposited data and metadata
7. Withdrawal/deletion interfaces

Fuller description of the work to be done follows. All components will be tested on samples to verify functionality.

1. User registration system

a) Define elements and build a system to describe and register users and communities, assign permission levels to and authenticate users.
b) Define and build the interfaces to enter and manage data about registered users and communities.

2. Digital objects (data and metadata) to be archived

a) Inventory the library's digital resources to identify format types, number of resources produced and acquired, the producers of resources and the rate at which resources are being produced and acquired. Include complex and multimedia

digital objects and web sites. Focus on the set of documents described in the Appendix: Proposed Content

b) Extend the inventory to university resources that have been brought to the library's attention, such as digital video of campus lectures and performances.

3. Metadata for management of digital objects

a) Define and implement a core set of metadata elements necessary for managing digital objects, including descriptive, structural and administrative (preservation, technical, rights and permissions) metadata. The metadata set should be aligned with standard metadata sets already developed by the digital library community (see the Library's Non-MARC Metadata Group. http://www.lib.uchicago.edu/staffweb/groups/metadata/links/static.html)
b) Specify how to package metadata from a variety of descriptive formats, e.g., DC, EAD, into MPEG-21, DIDL or METS, wrapper formats for metadata and data.

4. Deposit interfaces

a) Design and develop interfaces for the deposit of digital objects by users into the Archive.
b) Define and implement metadata that needs to be entered or that can be extracted.

5. Dissemination interfaces

a) Design and develop interfaces for the discovery and dissemination of digital objects by authorized users.
b) Define and implement the formats and metadata that will be disseminated.

6. Management of digital objects in the Archive

a) Define actions that will be taken on the digital object formats that will be deposited (e.g., virus, fixity and validity checks and metadata extraction).
b) Define how significant properties will be preserved.
c) Define how normalization will be done for different formats.
d) Document how the use of persistent identifiers, duplicate copies and migration will support long-term access.

7. Withdrawal/deletion interfaces

a) Design interfaces for withdraw/delete requests for items or classes of objects on an ad hoc or regular interval.
b) Define what metadata is recorded about the withdrawal/deletion.

Proposed Content

List of digital objects to consider in Phase 1 for digital archiving:

- Library web sites created by bibliographers and other staff to assist users in identifying and locating analog and digital resources—web sites consist of complex collections of still images, encoded text, databases, and scripts that underlie the interface

- Digitally reformatted versions of books and images from the library's collection (5,000 digital objects consisting of multi-file objects—50,000 digital files per year)

- Digital objects gathered and deposited into the SCRC Archives—these have certain restrictions on immediate use and must be migrated into the future so that they can be accessed in 30 or 50 years

- Library produced documents to support internal functions—often these exist only on the personal computer on which they were created or in attachments to email

- Digital metadata in a variety of non-MARC formats, EAD, Dublin Core, METS, and custom formats, provides for discovery of the library's collections in analog and digital formats and are digital objects themselves that must be archived

- External documents, discovered via the Internet, for which we want permanent access—we need to devise a way to download and maintain these digital objects in case the original publisher does not provide permanent access

- University events captured on audio and video that the university has asked the library to maintain permanently

- Digital data deposited in the Library by other units of the University.