THE UNIVERSITY OF CHICAGO

Report of the Digital Archiving Task Force

August 2004

Digital Archiving Task Force

Ron Thielen (Chair) *Assistant Director for Strategic Initiatives and Architecture, NSIT*
Bob Bartlett *Director of Enterprise Network Services and Network Security, NSIT*
Charles Blair *Co-Director, Digital Library Development Center, Regenstein Library*
Michael Fary *Associate Director of Data Administration, NSIT*
Chad Kainz *Senior Director of Academic Technologies, NSIT*
Benjamin Kessler *Director, Visual Resources Collection, Department of Art History*
Elisabeth Long *Co-Director, Digital Library Development Center, Regenstein Library*
Daniel Meyer *Associate Director and University Archivist, Special Collections Research Center, Regenstein Library*
Judith Nadler *Associate Director*, *Information Resources Management, Regenstein Library*

1. Executive Summary

Introduction

The Task Force was charged to consider the long-term implications of the growing percentage of University business conducted exclusively online and to recommend some near-term solutions that would improve current practices and address identified risks. Because of the potentially overwhelming size of the problem, the group was asked to focus its deliberations on identifying potential actions and associated costs for five cases:

- Electronic mail
- Web pages
- Administrative records (including student, financial, human-resources, and similar areas)
- Instructional materials (such as those in Chalk and electronic reserves)
- Research datasets

There are generally three reasons to preserve an institution's digital information:

- for legal and regulatory compliance
- to improve the operational effectiveness of the institution
- to preserve information with enduring value for historical purposes

The University of Chicago has substantial amounts of digital information that should be preserved for some or all of the above reasons. The lack of a digital data preservation strategy exposes the University to certain risks that due diligence requires us to address. This document presents some of those risks; describes the technical, political, and policy challenges to addressing them; and recommends certain actions that can be taken in the short, intermediate, and long term.

It is important to note that many organizations around the world are struggling with these same issues. The University of Chicago participates in working groups that are developing international standards for describing archival data. While we need not wait for these various efforts to bear fruit, our actions should be informed by their deliberations and directed so as to allow us to leverage the work that results from their efforts.

Summary of General Recommendations

Bitstream Preservation Program

A true archival program is a costly undertaking in terms of implementation effort, ongoing staffing, capital investment, and operational support. We recommend that the University initially invest in a program of *simple bitstream preservation* while committing to a long-term goal of active records management and, eventually, the establishment of a real digital archive.

Records Management Program

The University should establish an effective records management program to document its history, meet legal standards, support efficient and consistent administration, minimize the cost of records retention, and ensure long-term preservation of essential records. This records management program should mandate and authorize policies and practices for collecting and managing files, including:

- Establishing records management retention and scheduling policies with consistent application to similar or functionally parallel files across campus. Selection of materials for retention should reflect:

    - Legal, operational, intellectual, and cultural values
    - Need for near and long-term viability of files
    - Institutional responsibility
    - Institutional capacity
    - Depth of coverage commensurate with purpose for retention

- Establishing the necessary infrastructure. Infrastructure should be flexible in order to take advantage of technological advances, cost improvements, and staffing capacity. Existing infrastructure should be utilized where possible to contain cost and leverage in-house competencies.

- Assigning levels of preservation, considering:

    - Institutional values
    - Cost
    - Complexity
    - Functionality
    - Accessibility
    - Format

- Supporting methods of acquisition (capture) commensurate with type of archival material. These methods should simplify the burden on the end user supplying data and maximize the use of derived data.

- Using standardized metadata to manage data, including:

    - Metadata harvested from the source data
    - Metadata supplied at the time of archiving

- Establishing policy standards for archiving formats. The archive may commit to different levels of preservation based on such formats.

- Establishing policy for access. Present University policies governing access to its analog archives should be extended to include access to digital archives.

- Instituting an active media management process that includes:

  - Multiple copies
  - Storage environment
  - Accessing data regularly
  - Periodical refreshing of data
  - Migrating the archive to a newer technology base with some periodicity

- Establishing a funding model combining University budget and, where possible, external funding.

- Reviewing and reassessing records management polices and practices with some periodicity.

- Disseminating and ensuring compliance with policies and practices for the program.

Follow-on Group

The establishment of a Records Management Program and a digital archive will require a combination of technical solutions and policy decisions. A group should be formed to build upon the work done by this committee and consider in further detail the issues and needs surrounding the archiving of University material.

Summary of Specific, Short-term Recommendations

Infrastructure

- Extend existing tape capacity with addition of WORM (Write Once, Read Many) media, tape drives, and possibly additional robotic tape storage slots
- Add suitable disk capacity to "buffer" data being moved to tape
- Implement databases as appropriate to contain archival metadata
- Add servers to support the above
- Implement media preservation best practices, including examining the possibilities for additional environmentally controlled storage, possibly off-site
- With the exception of any costs for environmentally controlled storage space, the range of costs for these recommendations is between $120,000 and $200,000

E-mail

- Implement a mail server that will receive mail either automatically directed to it according to business logic rules or redirected from users to the mail archive

- Implement a commercial document archival and retrieval solution that will extract metadata from the mail being archived as well as any attachments and store the mail and attachments on a space-managed file system that migrates to and retrieves data from the tape archive
- Given the above recommended infrastructure additions, the initial incremental cost of archiving e-mail could range between $30,000 and $200,000

Web

- Implement a "web crawler" that will periodically create an "off-line" copy of the university's primary, openly accessible web sites. This web crawler would require a dedicated system on which to run.
- Store the off-line copies on a space-managed file system that migrates to and retrieves data from the tape archive
- Periodically create archival snapshots of the web servers' native file systems
- Given the above recommended infrastructure additions, the initial incremental cost of archiving web sites would be between $10,000 and $50,000, depending upon whether "shadow" web servers become necessary to offload the web crawler workload from production web servers.

Administrative Data

- Platform or application-specific backups or extracts of data should be moved to the managed storage infrastructure recommended above.
- Descriptive data, such as database schemas or COBOL copybooks, should be stored with the data extracts.
- Metadata describing the operational environment from which the data was extracted should also be stored.
- The costs to implement a preservation strategy for administrative data in this manner are entirely staff costs to develop the additional backup and extract processes.

Course Material

- Continue current preservation practices
- Establish a set of intellectual property policies and practices
- Establish a policy requiring centrally-hosted course content to be preserved in accordance with national and international standards and specifications for learning objects and instructional material
- Charge a working group to develop an academic content archival model
- Enhance the centralized authentication and authorization infrastructure to support the role and attribute based access requirements for archival content

Research Data

- Charge a working group to develop a specification for a University of Chicago academic data bundle (ADB) which will contain the data and metadata required to archive digital material in a useful manner and according to current and emerging international standards
- Establish a policy that defines requirements for an official University of Chicago research data archive (CRDA)
- Develop funding mechanisms to establish and maintain the CRDA
- Create a central, core data archive that addresses the CRDA requirements. This archive would be the reference model upon which other, distributed archives could be built and to which they could connect to form a federated research data archive.
- Enhance the centralized authentication and authorization infrastructure to support the role and attribute based access requirements for archival content

2. Recommendations

2.1 General Recommendations

These recommendations are intended to apply to the University at large, regardless of specific use case.

2.1.1 Recommendations for Preservation

Of the four preservation strategies, Backup, Simple Preservation, Records Management, and Permanent Archiving, the Group recommends that the University begin immediately by instituting a system for Simple Preservation for the materials identified in the charge. This will result in the initial creation of a so-called dark archive, or an archive that does not have a publicly accessible interface. (A management interface for the archive administrators will, however, be provided.)

The University should concurrently plan to follow Simple Preservation with a Records Management system, but such a system has as a prerequisite the development of an adequate access, authorization and authentication infrastructure.

2.1.2 Recommendations for Infrastructure

A certain amount of infrastructure will be required in order to implement any institutional digital preservation service. Our hope is to minimize the costs involved by leveraging existing infrastructure wherever possible. However, we will not be able to create an archive service by relying entirely upon already committed, or often over-committed, existing resources.

The minimal elements required are:

- on-line storage, i.e. disk
- off-line storage, e.g. tape
- a metadata management infrastructure consisting primarily of an appropriate database or databases and standard metadata management processes
- storage management software to manage the archival objects and media
- web site to front-end the service (documentation, registration, file submission, etc.)
- servers to implement the above
- environmentally controlled storage for long-term storage media

Storage options

Disk Storage

The type of disk storage required for an archival service must be reliable, available in large and inexpensive increments, and not necessarily of the highest performing type. A typical disk device suitable as part of an archival system has a 3-year acquisition and operational cost of under $50,000 and holds three terabytes of data. These costs decline year-to-year while capacities increase; however, the amount of data needing to be archived is also growing in step with cost/capacity improvements.

This is an area where technology is rapidly advancing. Several vendors are developing disk storage products specifically oriented toward archiving, primarily due to regulatory pressures from legislation such as Sarbanes-Oxley and HIPAA[1]. Our storage infrastructure needs to be flexible in order to take advantage of technological advances and cost improvements.

Tape Storage

While tape has been around for many years, it continues to have a role in the long-term storage of data. At some point, disk technologies may entirely supplant tape, but that is still some years away.

The advantages of tape are cost, portability, and the ability to create an unalterable copy of the data through WORM (Write Once, Read Many) technology. While disk and optical storage also have this WORM capability, the disk technology to do so is nascent and optical storage has other problems described below.

The disadvantage of tape is performance. The time to access a tape can range from several seconds for a tape in a robotic tape library to days for a tape stored at an off-site storage facility. Once the tape is placed in the tape drive, it may take several more minutes to position the tape to the file that is being retrieved.

Optical Storage

Optical storage (such as DVDs or CDs) has been the archival media of choice for many years. Unfortunately optical technology has not kept pace with the explosive growth of data. The largest optical media coming to market store approximately 27 GB, while 9GB is more typical. The advantage of optical media is a relatively long shelf-life if stored under proper environmental conditions.

[1] The Sarbanes-Oxley Act of 2002 established new rules for corporate governance among publicly-held companies. While the Act does not currently apply to non-public companies — including not-for-profit organizations — it establishes new or enhanced standards for corporate accountability. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) established national standards for health care transactions and addressed the security and privacy of health data.

Optical storage may be applicable in certain, limited circumstances. For example, relatively small datasets that need to be preserved in a portable, and platform independent format. However, we believe that the capacity limitations of optical media make it unlikely to have general applicability to our needs.

Storage Recommendations

Storage should be thought of as a hierarchy. Archived objects may be placed initially on disk but will be automatically migrated to less expensive, more portable tape or optical storage over time.

We recommend leveraging as much existing infrastructure as possible. For storage, this means using the Storage Area Network and NSIT recommended SAN disk subsystems. A disk storage pool of less than 3TB should be sufficient to support most archival data needs at this time. Objects occupying more than half of that space should be dealt with as special cases. The total cost over 3 years for 3TB of suitable disk storage is less than $50,000.

We can also use existing infrastructure for long-term storage through the newly acquired IBM 3584 tape robotic library. The IBM 3592 tape drives in this library support WORM tape cartridges. These cartridges currently cost approximately $100 each and can hold between 600GB and 900GB of data, depending upon how compressible the data is. We suggest acquiring an initial allocation of 200 WORM tape cartridges at a cost of approximately $20,000. We should also consider adding two more tape drives to the library at a cost of approximately $30,000. The library currently has available capacity; however, these recommendations will require that it be expanded sooner than it otherwise would have been. The cost for adding 400 more empty tape slots to the library is approximately $20,000.

We do not recommend introducing optical storage into the University's infrastructure at this time. For unique projects or applications where optical storage might be appropriate, we recommend considering the use outsourced media conversion services. This necessitates that guidelines be created that specify under what circumstances this would be appropriate.

Storage Management Software

There are at least two types of storage management software that may be necessary to implement an archival infrastructure. The first is generalized document archival and retrieval software. This sort of application stores archival objects, such as an e-mail message, along with the metadata associated with the object. The metadata is stored as indices within a database and the archival objects may be stored within a database or simply within files on a file system. The cost for this type of software starts at about $30,000.

The second kind of storage management software provides hierarchical storage and media asset management. It automates the movement of archived objects from disk to tape according to predetermined policies. We currently use IBM Tivoli Storage Manager (ITSM) for this function for backups. ITSM could also be used to support file and media management for archival files. The cost to license an additional ITSM server is under $2,000.

The disadvantage of using ITSM for archival support is that the data is stored in a proprietary format. An ITSM server is required in order to retrieve the data. This vendor dependence is not desirable in establishing an archive. However, ITSM could be used initially given the minimal license costs and the existing support structures within NSIT. We reservedly recommend using ITSM initially with the further recommendation that non-proprietary ways of performing the same functions be examined at some future time.

Database(s)

NSIT effectively has two database standard platforms that it supports Oracle and SQL Server. We certainly recommend that any database development required in-house be done using one of those standard platforms. We also recommend that commercial products with database components be given much greater consideration than those relying on other competing or proprietary platforms.

Servers

We do not view archival processes as having to perform at the same sort of levels as critical applications such as the library's Horizon system or student information systems, or course management system. Generally we anticipate that low-end servers may be employed to support much of the archival infrastructure.

Furthermore, we expect to be able to leverage existing servers and services to support some parts of an archival infrastructure. For example, Oracle databases could be housed within the general-purpose, NSIT Oracle service currently under discussion.

An additional ITSM server to support the media and hierarchical storage management functions would cost approximately $20,000, including 3 years of maintenance.

Environmentally Controlled Storage

NSIT-managed data center space can store media with short or medium-term storage requirements. For longer-term (anything more than 10 years), we should find storage space that has more restrictive environmental controls. The environment for long term storage of media is literally “cold storage.” The temperature and humidity would both be low enough that the environment would be unlikely to suitable for housing electronic machinery that is in normal use.

Having a second storage location at some distance from the NSIT data center would also have the advantage of providing storage for a second, disaster recovery copy of critical archival material. We recommend that use of suitable, long-term storage space for digital media be investigated.

2.2 Specific Recommendations by Use Case

The following recommendations pertain to the specific use cases upon which the Task Force was asked to deliberate: E-mail, Web, Administrative Records, Course Material, and Research Data

2.2.1 Recommendations for E-mail

Goals

There are a number of goals that should be considered for an e-mail archival system:

- The user should be able to easily insert messages into the archive.
- The archival service should be easily configured so that some subset of the population is identified as having all or most of their messages automatically archived.
- The archival service should be adjustable, preferably by the user, to identify messages matching a criterion take an action. This would permit a user to define which messages, if the automatic archive flag is set, will be stored into the archive.
- The user should be able to remove items from the archive if necessary.
- The archival solution should, if feasible, be part of a general archival solution and not a solution specific to e-mail.

Recommendations

The e-mail solution should not, unless absolutely necessary, be an independent archiving infrastructure. It should be part of a larger construct that addresses archiving for the wider range of issues. Most of the issues that must be solved for e-mail must similarly be solved also for the other issues. For instance, mail includes attachments of the same formats that must be handled when we address archiving the web, administrative records, or research. These solutions should be leveraged. Additionally, by having a unified solution, we can minimize personnel, training, and product costs.

Submitting E-Mail to the Archive

We propose that the primary method for submitting e-mail to the archive be by sending e-mail to an archiving mail server. Most users are already intimately familiar with their e-mail client and the basics of e-mail transactions. We can leverage this knowledge by installing an archiving e-mail server which will accept mail, store it, and from which it will be archived. As an intermediate solution, an e-mail server can be set up that will accept mail and store it into a mailbox. This mailbox can be regularly backed up and transferred to media that can be delivered to the archivist. To utilize this service, a recipient can “bounce” or “redirect” mail received messages to *user*@archive.uchicago.edu and they will be accepted and stored in a “dark archive.”

Also, a sender can carbon copy the archive address on any outgoing correspondence and it will similarly be stored in the archive.

Once a true archival solution is installed for e-mail, this server can be leveraged to forward mail to this receptacle for true archiving. Doing so should be transparent to the user and involve no change in user behavior.

Automation

Some automation can be implemented on behalf of the user to streamline archiving of e-mail messages. The NSIT mail server can recognize patterns in a message and act accordingly on the message. For example, the server can be configured to recognize a pattern and forward a copy of any message containing that pattern to the archiving server. This behavior can be utilized on messages either sent or received by campus users. The best place for a pattern to be placed is either in the e-mail address (*user*+archive@uchicago.edu or *user*@archive.uchicago.edu) or in the subject line ([ARCHIVE] placed on the subject line). We propose that the address be the place where the pattern be placed. A "personality" can be set in the e-mail client that defines the address or reply-to address according to an agreed upon pattern; any mail sent using that personality will be recognized by the NSIT mail server and be copied to the archive. Likewise, when the recipient replies to a message sent using the archive personality, the response would automatically be recognized and archived as well.

Additionally, an object could be defined in the NSIT LDAP (Lightweight Directory Access Protocol) directory server to permit a user to set mail archiving on all messages being sent or received by the user's address. This flag could be set using the current CNet account editing page to set this value. A possible extension to this concept would be to design a web page that would permit a user to set and delete more granular filtering rules for the matching patterns recognized by the mail server. For instance, users could define a rule that would set any message to or from them that also has a particular address in the header – either To:, From:, or Cc: -- to be copied o the archive.

2.2.2 Recommendations for the Web

Goals

There are a number of goals that should be considered for the web content archiving system:

- Web site administrators should have a mechanism for registering web sites for archive.
- Critical areas should be designated as such and should be regularly archived.
- The user should be able to submit descriptive metadata for a site and have that retained as part of the archive.
- The archival solution should be part of a general archival solution.

Short-Term Recommendations

We propose a dual-path approach to archiving over the short-term. First, we propose the installation of a special "spidering" web archiving machine. The function of this machine will be to use a special program known as a "web crawler" to traverse designated web sites and download the contents of those web sites for archiving. The crawling program also has the capability to rewrite links so that they point to files in the local file repository rather than on the remote websites. Doing so will allow the archivist to restore a local copy of the web content of any site or any subsection of a site from archive. The limitations of this are that the crawler can only access public areas of any web site; only unless it is given special privileges can it access restricted areas. Such privileges should be applied to areas that are designated important for archival purposes. As the access permissions are bypassed by using crawling software, access restrictions will have to be noted elsewhere for the archivist.

Second, we propose leveraging the backup strategies used by the systems administration staff of the web servers to generate "dark archives" of web content. Regular special backups can capture the state of the web content and can be turned over to the archivist. When needed, the web content can be restored to a system that is customized to run a web server and present the pages to the archivist or interested party. This procedure should only be used for data that cannot be retrieved by the previous crawling method.

Neither of these interim methods is future-proof. As various data formats, archiving formats, and web programming tools evolve and change, the data contained in the dark archive will become obsolete.

Long-Term Recommendations

Our long-term goals are to have a web archiving infrastructure that is not separate from the rest of the archiving infrastructure. This infrastructure should address the following needs:

- It should maintain and allow extraction of time-ordered versions of web sites. Ideally this would be accomplished by maintaining version control over all web source files which would be archived. However, this could be accomplished by taking snapshots of any archived web sites.
- The archiving infrastructure should maintain an indexed database that permits easy and quick searching of the archive to identify sites or documents of interest.
- The archiving infrastructure should permit role-based access to its contents and the authentication method should be based upon LDAP (Lightweight Directory Access Protocol). Authorization levels for access to the archive should be partitioned so as to restrict access to specific areas within the archive.
- The archiving infrastructure should incorporate tools that present data contained within the archive in a useful way. This capability will need to be upgraded as data formats evolve and change.

There is no product or collection of products that presents itself to the committee at this time as a solution. We recommend that a follow-on committee be convened to investigate the solution to these requirements and propose it to the University.

2.2.3 Recommendations for Administrative Records

Archiving extractions should occur from the administrative system data stores on regular intervals. These intervals could be monthly, quarterly, semi annually, or annually. In many cases the extractions should be completed after a normal business cycle update occurs (e.g. monthly closing, year end closing, begin/end of a quarter, etc.).

Short term recommendations

For each of the data stores:

- a vendor utility backup or exported copy of the files or database should be kept with a long term expiration
- a current schema or file layout of the data store at the time of the extraction should be included
- metadata regarding the hardware/software environment should be kept

For data not residing on the mainframe, these archival objects can be moved to the recommended storage infrastructure through the use of existing backup and production automation agents.

For mainframe data stores, data can be archived directly from the virtual tape subsystem that Data Center Services is planning to implement later in 2004. The mainframe virtual tape system has the ability to store data directly into the managed storage infrastructure.

Longer-term recommendations

For each of the data stores:

- each file/table should be extracted and written to a file, with values stored as plain text
- any binary, encoded, or computational fields should be converted to plain text

For each of the data stores, the following should be included with the archive as a bare minimum:

- current schema or file layout of the data store at the time of the extraction
- data dictionary of all data elements being extracted
- lookup or reference tables or files, extracted and written to a file with values stored as plain text

Note: any files/tables which contain data which is necessary for the functioning of an application accessing the tables should be excluded from the archive. These might include tables/files which contain screen names, screen formatting data, stored procedures, subprograms, configuration information, etc.

Longest-term recommendations

For each of the data stores, the following should be included with the archive:

- scripts should be developed to associate related records and extract them in a referential format. Included in the extracts should be populated values from any lookup or reference files or tables.
- a complete set of dated lookup or reference files should be provided
- a taxonomy of access rules for each data store by user and organizational unit
- for data stores where graphical data is stored elsewhere (e.g. on a file server), the graphical images (timecards, invoices, etc.) should be retrieved and stored with sufficient relevant data from the source data store to enable identification of the graphical data

2.2.4 Recommendations for Course Material

Because the current state of course content is mainly centralized and stored on Chalk, short-term archival issues regarding course content become largely matters of policy rather than technical implementation. In the long term, the recommendations turn toward automation on a technical level.

Short Term

- Continue archiving Chalk course content using current procedures including off-site storage and transfer to DVD.

- Establish a set of related intellectual property policies regarding course content that address the University's investment in course content, the faculty's rights to the content, and the ownership of student-created material that contributes to and is incorporated within the course.

    - The University should explore a limited (three-year), non-exclusive license to access and reuse course material to minimize risk to the institution in the event a faculty member is unable to continue teaching a specific course. A three-year window is a reasonable timeframe to develop a replacement course using new content and a different approach. After the expiration of the license, course content will be retained for archival purposes as part of the University's institutional record of the course offering.
    - Faculty may choose to share course elements (learning objects) with other faculty. The policy should define learning object submission criteria,

    including the perpetual non-exclusive licensing of the content by the institution.
  - Students regularly contribute to course content, whether through discussion or academic activities. Therefore, a policy should define what is protected student information within a course, what is considered the student's intellectual property, and what constitutes contribution of that intellectual property to the institution and its archive. The latter can be defined by a formal non-exclusive use license similar to the sharing of learning objects by faculty mentioned above.

- Establish a brief policy that requires centrally-hosted course content to be archived in accordance with established and emerging national and international standards and specifications for learning objects and instructional material.

  - The policy can state that course content hosted on local servers and personal home pages will not be archived as instructional content, but may be included as part of archival procedures related to Web content (outlined elsewhere in this document) if hosted on institutional servers such as www.uchicago.edu, home.uchicago.edu, or www.lib.uchicago.edu.
  - Course content falling outside of these areas may still be archived, but as an academic data bundle through the proposed depository model for research data (outlined elsewhere in this document).

- Charge a working group to develop an academic content archival model for learning objects, instructional content, and related digital media that incorporates, once developed, the academic data bundle (ADB) specification used for research data.

- Enhance the current centralized authentication and authorization infrastructure to support distributed role-based and attribute-based access to content using established and emerging authentication and authorization technologies used among research universities (also related to Research Data).

Long Term

- Establish a policy defining course content as eligible for being archived alongside research content.

- Automate the creation of an academic data bundle for each course hosted on Chalk, and archive the bundle within the University of Chicago Research Data Archive (CRDA) as defined in the Recommendations for Research Data section of this document.
- Discontinue the archiving of course content to DVD and transfer existing content as academic data bundles into the CRDA.

- Develop a central NSIT/Library digital object and asset repository (DOAR) that fully implements the extended federated CRDA (outlined in Recommendations in Research Data), thus acting as an archive for academic digital media assets and learning objects that may fall outside of research and instructional materials (also related to Research Data).

- Migrate online academic data- and digital asset-related projects developed by the Library and NSIT to a new architecture that depends on DOAR for asset management and benefits from transparent integration with the institutional archive (also related to Research Data).

    - Once a project model is established, open the architecture to other groups on campus including units interested in developing learning objects and research support operations.

2.2.5 Recommendations for Research Data

Because research data can encompass virtually any data type (ascii text, unicode text, integer, floating point, double floating point, binary, etc.) in any number of forms (data sets, applications, scripts, etc.) of virtually any size (megabytes upwards and beyond terabytes), we recommend that the archiving protocol for research data should follow a depository model where the material is stored as a single bundle with appropriate metadata describing the research program, bundle characteristics, reuse requirements, digital rights information, etc. In addition, we feel that content stored within the archive should be subject to the current and future guidelines and policies regarding intellectual property and should not be treated as a separate entity.

Assuming that the University chooses to implement a depository model, we recommend that a policy be established that states that it is the University's responsibility to ensure that the bundle can be opened, its component parts restored, and the data integrity validated. In response to the policy, the University would establish appropriate data integrity validation procedures to ensure the content does not degrade over time as technology advances, and ensure the ongoing management and support of the archive. It would be the responsibility of the requesting user or entity to provide storage for the unbundled content that meets the reuse and digital rights requirements stated in the bundle. If the data integrity of the bundle is determined to be valid, the University's responsibility ends as the ability to read and/or reuse the data becomes the responsibility of the requesting user or entity.

Because of the nature of contemporary research, which is often inter-disciplinary and inter-institutional, a single central data archive may not be feasible or attainable. Therefore, the definition of a research data archive may extend to a managed, distributed, and federated collection of individual archives bound by a common strict criteria of interoperability conditions, management procedures, and institutional policy requirements; any of which not met may invalidate an individual federated archive from formally being included in the campus archive and being afforded the legal protection as an institutional archive, as well as the administrative support of the University.

Short Term

- Charge a working group to develop and publish a specification for a University of Chicago academic data bundle (ADB). The ADB will minimally contain the digital material to be archived, appropriate metadata describing the research program, bundle characteristics, reuse requirements, and digital rights information, and be an application profile of documented open standards and specifications. Given the long-term nature of archived content, the working group should also develop an approved institutional process for amending the ADB that is able to address future needs while maintaining the viability of legacy ADB-based content.

- Establish a policy that defines detailed and specific requirements for an official University of Chicago research data archive (CRDA) that incorporates the ADB specification and encompasses contributions and requirements from the Office of Research Administration, the divisional Institutional Review Boards, the Library, NSIT, and other appropriate units on campus. The policy should include auditing and tracking rules.

- Develop a funding mechanism to establish and maintain the CRDA. Funding could come from an appropriate allocation out of the calculated grant overhead.

- Create a central core data archive that addresses CRDA policy. This archive would be the core archive of the future federated archive model. It would become the reference archive to which all federated archives would bind, comply, and conform.

    - Funding for the core archive would come from a number of sources, but in terms of research, funding should have a calculated base to cover ongoing management costs and, in addition to the base, be proportional to the amount of research content archived in respect to the average cost of physical storage amortized over time.

- Enhance the current centralized authentication and authorization infrastructure to support distributed role-based and attribute-based access to content using established and emerging authentication and authorization technologies used among research universities (also related to Course Content).

- Promote and encourage the use of the CRDA in grant-funded research.

- Explore long-term storage issues and recommend a minimum length of time to maintain data in the archive.

Long Term

- Charge a working group to develop technical interoperability criteria for a federated version of CRDA. The working group should include representation from across campus as well as major research entities (including Argonne, for example) and encompass the state of technology across higher-education research universities and research-specific organizations.

- Establish a sustainable campus digital rights management infrastructure that includes as part of its architecture, the academic data bundle (ADB) specification and needs of the CRDA.

- Extend the CRDA to a distributed and federated archive model built upon the technical interoperability criteria for a federated data archive, the digital rights management infrastructure, and enhanced authentication and authorization services.

- Develop a central NSIT/Library digital object and asset repository (DOAR) that fully implements the extended federated CRDA, thus acting as an archive for academic digital media assets and learning objects that may fall between research and instructional materials (also related to Course Content).

- Migrate online academic data- and digital asset-related projects developed by the Library and NSIT to a new architecture that depends on DOAR for asset management and benefits from transparent integration with the institutional archive (also related to Course Content).

    - Once a project model is established, open the architecture to other groups on campus including units interested in developing learning objects and research support operations.

- Amend the CRDA policy to include a defined minimum length of time to maintain data in the archive.

- Require as official University policy the use of the CRDA in grant-funded research.

3. General Background Issues

3.1 Introduction

The recommendations of the Digital Archiving Task Force are based upon the examination of technological and administrative issues affecting the management of electronic data. These issues will first be discussed in their most global sense as they affect the University at-large. This will be followed by discussion of the specific use cases examined by the Task Force.

Risks and Context

The University is not alone in facing the potential loss of its recorded operations and activities due to the fugitive nature of digital materials. The critical problem of digital preservation has been recognized at the national level in the charge to the Library of Congress to develop a National Digital Information Infrastructure for Preservation (NDIIP) and by a variety of conferences and projects sponsored by organizations such as the Coalition for Networked Information (CNI) and the Digital Library Federation (DLF). But while we should track national efforts, initial action cannot wait upon future national solutions. Digital information presents challenges that are qualitatively different from those of paper. Storage media have shorter life spans, data is more easily altered without trace, and access requires a combination of hardware and software which quickly become obsolete. Ensuring long-term access to digital materials requires the establishment of new life-cycle management strategies which must begin far earlier in the life of the object than is required within the paper environment.

The concern over potential loss of digital information stems from three key values that the community places on its digital output. Items may have legal value, operational value, intellectual/cultural value, or any combination of the three, and the risks associated with data loss vary accordingly. Items admitting of legal value need to be maintained in order to comply with provisions of federal, state, or local law and with institutional legal commitments, or in order to maintain and protect the institution's legal position. Laws such as the Family Educational Rights and Privacy Act (FERPA) or the USA Patriot Act affect what records should or shouldn't be kept, and conditions on grants received require the University to maintain and make available certain data for specified periods of time.

Materials with operational value need to be maintained in order to assure the efficient administrative functioning of the University. Records such as University policies, committee activities, administrative correspondence, etc. can be of ongoing usefulness to administrative operations and for maintaining continuity over time. Materials with intellectual/cultural value must be preserved over the long-term for use in the teaching and research activities supported by the University. Such materials may include items generated specifically for teaching and research, but can also include materials whose initial value is legal or operational but which gain in intellectual/cultural value as time passes. Thus the need to act quickly to safeguard digital materials being produced by the University is in response to the legal and operational risks that could result from loss of data, as well as from the institution's commitment to secure its own intellectual output and cultural significance.

Focusing on the last goal, that of collecting an institution's intellectual output, many universities have begun building institutional repositories to house faculty papers and, in some cases, to support new methods of scholarly communication. DSpace (MIT) and ePrints (University of Southampton) software are examples of the results of two such initiatives which have been made available to other institutions for use. Less has been done by way of supporting the broader range of needs, outlined above, for collecting and maintaining the wide variety of digital materials of potential interest to an institution. Nor are there any mature systems for managing the archival functions necessary for safekeeping and providing access to digital files over time.

Preservation strategies

Several general issues pertain to the collection and storage of any type of digital file, with additional case-specific issues applying to the particular types of records.

Devising a strategy for collecting and storing files requires the development of scope statements outlining the specific types of files to be collected and the reasons for doing so. Various technical solutions might address different goals, so final decisions must be based on the articulated value(s) of the records and the purpose for managing them over time. For instance, archiving the University's website in order to capture a prospective student's general experience might demand a front-end approach that could be scheduled at regular intervals every few months. Archiving the University's website in order to capture the policies it contains for legal and operational purposes might demand a more frequent and more ad hoc, event-based solution in order to capture every revision as it occurs.

At the same time, it is important to realize that an item's value changes over time, as does the frequency of its use. For instance, records that start out having high operational value and which are frequently used may, over time, develop intellectual/cultural value and/or become less commonly consulted. Solutions should account for all eventual uses whenever possible. Additionally, not all records will need to be maintained over the long-term. In fact, it is assumed that there will be a gradual winnowing of materials stored, with larger amounts of data captured and maintained for near-term use, and only a selective set being given the higher level of care required to maintain accessibility over the long term.

Four different levels of preservation can be applied to a record including *backup, simple preservation, records management,* and *full archiving*. Each stage varies in its cost, functionality, selectivity, and accessibility, with increasing technical and administrative complexity at each level.

Backup

Backups are necessary as part of an archiving strategy, because backups allow for the recovery of data in the case of unintentional modification or deletion. However, backups alone are not sufficient for archiving data because they are intended to serve only as short-lived copies of the physical bitstreams.

Simple Preservation

The goal of simple preservation is to preserve bitstreams and keep them unchanged over time. It includes a backup strategy, fixity checking and media refreshment in order to prevent data loss or corruption; but while it ensures that that data so preserved can be read in the years to come, it does not ensure that they can be interpreted.

Records Management

The third level of preservation provides active management of records, which implies a higher degree of selectivity, increasing amounts of associated metadata, commitment to migration of file format in order to preserve interpretability, and provision for an access system. Records management tracks the changes to a file throughout its life cycle, managing legal status, accessibility, retention requirements, etc. over time. A records management program applies consistent maintenance, retention, and disposal procedures and supports the University's ongoing operations.

Archiving

The fourth level of preservation is long-term archiving in order to permanently preserve the intellectual/cultural value of items. Only a subset of materials in a records management program will require permanent archiving. Permanent archiving requires a full complement of bibliographic and administrative metadata in order to ensure accurate migration and representation of data, and to provide on-going accessibility.

3.2 University Implementation Issues

3.2.1 Policies

Preservation of digital materials will require the establishment of a variety of policies. Some will be based on the intellectual content of the materials (e.g., what materials should be collected, who should get access to which records and when, etc.), while others will be based on technical issues (e.g., what file formats can be preserved) or on general program goals (e.g., what constitutes an acceptable archiving system). A central model and infrastructure for setting policies and defining functional system requirements would provide the University with the assurance necessary that any such system is addressing pertinent legal and operational requirements.

Authorization (Institutional Mandate)

The University of Chicago currently has no institutional mandates or policies for records management or archiving. The Board of Trustees has not taken any action to mandate the preservation of any institutional files, whether in paper, analog, or electronic form. The University Statutes and Bylaws do not authorize any University officers or staff to preserve or manage archival records.

In 1946, the University of Chicago Library, in accordance with its mission to support teaching and research, agreed to accept responsibility for selecting, managing, and preserving institutional records of the University with long-term historical value. The current size of the University Archives collections is 25,500 linear feet, the equivalent of 25.5 million individual pages of documents spanning the historical period from the 1850s to the present. Institutional records in paper form that are collected and managed by the Archives include: minutes of the Board of Trustees; files of the President, Provost, Vice Presidents, and other University officers; Registrar's Office transcripts; administrative files of divisions, schools, departments, committees, and other institutional units; and official publications, including reports, course announcements, directories, and newsletters.

At the level of the Trustees, President, and Central Administration, the University Archives maintains one of the most complete sets of administrative records of any American private research university; below that level, among divisions, schools, and departments, the amount of material retained and archivally preserved for future use has been dependent entirely on the varying administrative practices over time of successive individual deans, chairs, and office managers.

In order to create a systematic records management and archiving program, the University needs to develop an institution-wide mandate for the preservation of essential legal, operational, and intellectual/cultural materials. This mandate will

- Define the mission and purpose of a records management and archiving program
- Identify the functions the program will perform

- Identify the officers and staff responsible for its implementation
- Outline the processes to be followed in developing policies and management practices

Policy (Archival Content)

Scope and Content

The University's policies for records management and archiving should have a breadth and consistency that will make it possible to apply them across all forms of digital information managed by the University, including electronic mail, Web pages, administrative records (student, financial, human-resources, and similar areas), instructional materials (such as those in Chalk and electronic reserves), and research datasets

Records Management and Records Schedules

Systematic programs for archiving of institutional records are typically shaped by a records management policy. Standards and practices for institutional records management have been developed and promulgated by a number of professional organizations, among which the largest and most comprehensive is ARMA, the Association of Records Managers and Administrators International http://www.arma.org/).

ARMA and its subsidiary units have created recommended policies for records management and administration, records scheduling and retention, legal risk management, government regulation compliance, privacy and rights management, and management of all forms of electronic records including e-mail. ARMA continues to revised and extend its recommended polices as new technologies are adopted for creating, storing, and disseminating institutional records. Training and certification of records management professionals is governed by the ICRM, the Institute of Certified Records Managers (http://www.icrm.org/).

A well-planned University digital records management program should be developed in accordance with the recommended policies and standards of ARMA, ICRM, and other professional organizations and should include:

- development of policies and procedures before the creation of records takes place
- a controlled system of formats and metadata for records that are created
- systematic processes for regular capture, transfer, and preservation of records on a defined cycle ("records scheduling")

Records management has clear advantages for a digital archiving program:

- records are created in formats and with metadata that supports capture/transfer

- documents are created and identified in a more deliberate manner that helps facilitate later file sorting or hierarchical arrangement in file directories
- extraneous or less significant files are more readily identified for later stages of archival analysis and ultimate decisions about retention/disposition/preservation

Archiving

Records management provides an effective and cost-efficient means to identify essential records and assure their maintenance through the information life cycle. Archiving assures that the most significant materials retained through records management, those with enduring institutional and historical value, will be selected and preserved on an indefinite or permanent basis.

The principal organization for the development of recommended archiving policies and practices in the United States and Canada is SAA, the Society of American Archivists (http://www.archivists.org). SAA has taken a leading role in creating and extending standards for the whole range of activities performed by professional archivists: acquisition, appraisal, description, rights management, and preservation. SAA is also an important venue for development of recommended policies and practices for all forms of digital archival materials: administrative records, data sets, electronic mail, educational course materials, oral history interviews, photographs, audio and video files, and architectural and engineering drawings and presentation materials.

The University's policies and practices for archiving, like those for records management, should conform to the recommended standards of recognized professional organizations. Drawing on the programs of SAA, ARMA, and ICRM will also make possible a more efficient and cost-effective implementation of the University's records management and archiving policies and infrastructure.

Selection and Appraisal

Selection of material designated for preservation involves analysis ("appraisal") of the functional processes of an institution and an evaluation of the significance of different record or file types to document these functions. The purpose of the appraisal is to ensure that functions identified as significant are documented at an appropriate level of detail and completeness, and to avoid retaining materials that do not meet these criteria.

Appraisal is not undertaken to evaluate material on a file-by-file or item-level basis. Rather, whole classes or types ("series") of records are evaluated, and decisions about retention and preservation are made about the entire class or type as a whole.

Records scheduling decisions (how long to retain material) are also made on the basis of a review of an entire class of material. Records schedules are not determined for an individual record or file.

Management (Repository Operation)

The management of a digital archive repository raises a series of issues that will need to be addressed from an institutional perspective:

Organizational Structure

- How large an organization is required
- How dispersed its components may be across the current University structure
- Advantages and disadvantages of centralized management

Funding and Budget

- Amount required for startup and continuation of digital archiving program
- Advantages of centralized funding vs. dispersed or cost-recovery models

Repository Location and Operation

- Equipment and space requirements
- Advantages and disadvantages of centralized servers and file management vs. dispersed facilities
- Potential physical locations of archival repository whether established on centralized or dispersed model

Security, Restrictions, and Access

- Controlled physical access to repository space
- Controlled access to file content for file managers, programmers, analysts
- Controlled access to restricted file content for internal use by initial creators of files or their authorized representatives during a restriction period
- Controlled access to restricted file content by the University community or by the general public at the conclusion of a restriction period
- Controlled access to file content that is non-restricted at the time it is added to the archive

The purpose of an access policy is to provide open access to files at the earliest possible date following their creation. Restrictions exist to ensure that confidential or sensitive material is not disclosed in advance of set deadlines for disclosure.

Files from a digital resource like the University Web (most of which is open to public viewing and use) will have a different pattern of restriction and access than a digital resource that has confidential or proprietary content.

Current archives access policies govern the use of official University records by faculty, students, staff, and other researchers. The policies cover two principal types of records:

- Administrative records of the central administration, including the Board of Trustees, its committees and sub-committees; the Office of the President of the University; offices of Vice-Presidents; and other administrative offices such as the Provost, Comptroller, Treasurer, and Registrar
- Administrative records created by the University's professional schools, graduate divisions, academic departments, the College, committees, centers, and other formally constituted units of the University.

Prepared in consultation with the Secretary of the Board of Trustees, archives access policies were first formulated in the early 1960s, revised in the early 1980s, and revised most recently in 2002. These policies are made available to researchers on the Special Collections web site: http://www.lib.uchicago.edu/e/spcl/recordsaccess.html

User access for the archival repository would be most effectively managed through a systematic set of chronological access periods, beginning from the date of creation of the archived file. The sequence below sketches in rough form how one possible sequence, based on current University Archives restriction and access policies, might operate:

- *Near-Term*: Digital files selected for Near-Term status would be held in their original fixed form for a period of 10 years from the time of their creation.

- *Near-Term Read and Copy Access*: Granted to the Dean, Chair, or Director (or designate), the relevant Assistant/Associate Dean, Chair, or Director (or designate), and the Creator of the file (if still retaining the same University position as when the file was created). Access is defined as read-and-copy access only; no file content in the Near-Term period could be altered, edited, or amended.

- *Intermediate Term*: Digital files would be selected from the Near-Term class and moved to the Intermediate Term status; they would be held in their original fixed form for the period extending from 10 years to 30 years from the time of their creation. Files containing budget and appointment (personnel) information would be held in their original fixed form for the period extending from 10 years to 50 years after creation.

- *Intermediate Term Read and Copy Access*: Access would be granted to the Dean, Chair, or Director (or designate), the relevant Assistant Director (or designate), and the Creator/Chair (if still retaining the same University position as when the file was created). Access is defined as read-and-copy access only; no file content in the Intermediate Term could be altered, edited, or amended.

- *Long-Term*: The third stage pertains to files in the Long-Term period (i.e., Preserved and Open Long-Term or Permanently). Digital files selected from the Intermediate Term to be moved into the Long-Term status would be held in their original fixed form indefinitely beginning from the point 30 years after the time of their creation. Files containing budget and appointment (personnel) information would be held in their original fixed form indefinitely beginning from the point 50 years after the time of their creation.

- *Long-Term Read and Copy Access*: Access during the Long-Term period (i.e., Preserved and Open Long-Term or Permanently) would be granted without any restriction. Access is defined as read-and-copy access only; no file content in the Long-Term class could be altered, edited, or amended.

File Ownership and Access Control: Repositories vs. Archives

The technological issues surrounding management of access rights may be simplified by policy decisions, primarily regarding the purpose of the archive. As described previously, when material is deposited in an archive it is so done for legal, institutional, or historical purposes.

The reason for establishing an archive is one of the factors that distinguish an archive from a repository. One may well contribute data to a repository with for the purpose of sharing it amongst colleagues, but that is not necessarily a reason for archiving material. A repository may play a role in constructing an archive, but it is not an archive.

Similarly, it is our belief that an archive is <u>not</u> intended to serve the same purposes as a backup facility. Therefore, it should not be a requirement that control over access to archival data be implemented or managed in the same manner as access control for backup data. Primarily this means that the original "owner" of the data, in the sense that the originating systems view ownership, need not automatically be given access to the archival data. In the long-term, initial ownership of the data becomes less relevant due to changing roles and responsibilities as well as life events.

3.2.2 Communications

In order for any digital preservation effort to succeed at The University, some "marketing" and communications effort will be required. These efforts should clearly communicate:

- The goals of the service
- Any relevant policy, such as appropriate use of the service
- The technical requirements for using the service
- The costs (even if they are not recharged[2])
- The limitations
- Directions for use
- Frequently-Asked-Questions and pointers to technical support

Clearly the world-wide web is the current vehicle of choice for delivering such information. Accordingly, a web site for digital preservation will be required.

[2] We believe that giving the potential users a feeling for the cost of providing such a service, even if they are not charged for it, will help to induce proper behavior.

Furthermore, some support from the NSIT help line and training organizations may be needed, but their requirements will have to be defined at a step closer to an actual implementation project.

3.2.3 Costs and Funding

Even in the most simplified scheme of bit-stream preservation, there will be both capital and ongoing costs. Development of a more mature system for records management and digital archiving will require even further funding. A integral element in any program for long-term maintenance of the University's digital files is a commitment to ongoing funding. Any file management strategy will fail, no matter how well-designed and built, if funding to support the replacement of outdated hardware, the migration from obsolete formats and software, and the staffing necessary to perform archival functions, is not maintained.

Phases

Developing a system for managing the long-term maintenance of digital files may include three phases: startup, pilot, and on-going. Startup and pilot phases may be able to attract external funding from granting agencies interested in solutions to the critical problems we face with digital files. On-going operation costs, however, will require a more stable and sustainable funding model that ensures support for at least a minimum-level of operations.

Staffing

As previously mentioned, we are recommending a strategy of simple preservation as an initial step that must be taken in the immediate future to prevent the loss of essential data.

Beyond this first step, which we believe is urgently required, we recognize that additional levels of preservation as outlined in this report will bring added financial cost and administrative complexity. The largest component of cost for a records management and digital archiving program will unquestionably be staff and supervision, both for the initial implementation and for all ongoing operations. Initial implementation will require the creation of a number of new dedicated staff positions, but the largest staff costs will be incurred through ongoing operations as the size and scope of the digital archive continues to increase.

Despite these expected costs, we nevertheless believe that the University is confronting a situation of growing legal vulnerability and potential loss of essential information of institutional and historical value if it does not take steps to institute an effective records management and archiving program. It is our recommendation that the full development of such a program, with all necessary additional staff positions and technical support, should be considered a matter of immediate and growing urgency and that the administrative planning and decision making required to create it should begin soon as possible.

3.3 Technical Implementation Issues

3.3.1 Intellectual Property and Related Issues

One strength of a research university lies in the knowledge and ideas it generates. The value of the institution is largely determined by its contribution to society, and as such, its intellectual capital becomes a significant asset. When one considers the creation of an archive that could house the administrative, social, and scholarly digital content of the institution, the issue of intellectual property must be addressed.

Who owns it?

In recent years, questions have arisen regarding who "owns" the intellectual property generated on a university campus, less so at the individual faculty researcher level where policies and guidelines exist, but more within the realm of undergraduate and graduate scholarship that crosses the boundary of coursework and research. There have been numerous examples of disputes between graduate students and faculty over experimental results. Such disputes are not entirely uncommon as that tends to be the nature of scholarship, but more recently these disputes have triggered researchers to claim intellectual property rights over experimental data, thus preventing graduate students from publishing dissenting articles based on the acquired information. Informal discussions among academic computing staff have revealed a growing trend of undergraduates making intellectual property claims over materials submitted online as part of course-related activities. Some have suggested that within MBA programs, it is becoming more common for students to require faculty to sign non-disclosure agreements to protect student ideas that may be presented as part of coursework. These trends raise serious concerns regarding the nature of contemporary scholarship and may, in the end, limit the flow of ideas that is critical to the growth, future, and value of a research university.

A case in point

The above situations represent rather clear one-to-one cases. However, when all cases come together into a single project, the intellectual property issues become extremely murky. To find an example, one need only turn to our own campus. In June 2004, the Board of Computing Activities & Services recommended funding the *HyperAtlas* project proposed by T.J. Mitchell as part of the Provost's Program for Academic Technology Innovation (ATI). This project is a particularly complex intellectual property problem. The grossly simplified (and understood thus far) background leading up to the ATI proposal and funding is as follows:

1. T.J. Mitchell (Humanities) offered a course on media theory to undergraduate and graduate students.
2. An undergraduate proposed a three-dimensional model to visualize the intersection and relationship between theoretical concepts and media.

3. Several graduate students explored the technological possibilities of creating an interactive environment for the model.
4. T.J. Mitchell, the students, and Lec Maj (Humanities Research Computing) approached NSIT for ideas and assistance.
5. A project team of undergraduate and graduate students, T.J. Mitchell, and Lec Maj is formed.
6. The project team envisions a model where multiple theorists, both on and off campus, can contribute to the project.
7. NSIT connected the project team with Jonathan Silverstein (BSD) for advice and direction on scientific visualization technology and techniques that could apply to the project.
8. Armed with that information, T.J. Mitchell submitted an ATI proposal.
9. BCAS recommended funding the proposal.
10. The Provost approved institutional funding of the project.

From a scholarly perspective, this project represents one of the many strengths of the University of Chicago. However, if one raised the question of "who owns what" in terms of intellectual property, the answer today is unclear. The project is based on a faculty-developed course, an undergraduate visualization idea, graduate student contribution both intellectually and technically, a different faculty member's research, and institutional contribution in time and money (NSIT and ATI). If the project succeeds, one needs to layer in the contribution of ideas of scholars from around the world. When the project is "complete" and stored within an archive, who "owns" the intellectual property at that point?

Policy options

Complex projects like *HyperAtlas* will become more common as collaborative learning models are adopted more widely on campus and begin to influence ongoing research. The University, then, needs to determine its intellectual property role and position in such situations. The University can:

- do nothing, or
- actively protect individual intellectual property rights down to the student level and develop policies and guidelines to protect individual rights and limit institutional claims, or
- adopt a position of intellectual property as being key to scholarship, thus extending institutional IP interests across all members of the community to promote open scholarship.

If the latter "open scholarship" position is adopted, material stored within an archive becomes an institutional asset that is not unlike the materials stored within the library, available to future scholars without fear of licensing and potential litigation. Developing such a policy or set of policies will be difficult at best and may involve new notions of licensing heretofore unseen on campus at the student level. However, the long-term implications will extend well beyond archiving material, and may serve to benefit the

University by reinforcing its institutional values around scholarship and the free exchange of ideas.

3.3.2 Access Rights

Any system for maintaining University-produced data will need to be able to manage access rights, since many materials will need to be restricted at least for some initial time-period. Restrictions policies will generally be role-based though the system will need to be able map roles to individual users. For example, President Randel may have the right to look at particular records, but that right is associated with his position rather than an inherent personal right. Not only do people change their roles within the University, but the roles themselves change over time as the University changes its structures, so there will need to be a mechanism for transferring rights as roles change.

Even though the most practical immediate solution may be to build a "dark archive" which safeguards files but does not allow access, depositor information will need to be recorded from the start so that future access systems will be able to act upon it. This implies the need for a metadata scheme (with elements and their range of values) and a mechanism for recording the information.

3.3.3 “Push” vs. “Pull”

Archival acquisition methods can be divided into two fundamental models, described here as “push” and “pull”:

- A “push” model consists of the active contribution of data through either manual or automated processes. A manual archiving process requires someone with authority over the data to take some positive action to cause the data to be sent to the archival system. An automated process would replicate this authoritative procedure by being integrated into a workflow for creating or managing the data.

- A “pull” model, on the other hand, entails a passive capture of the data through asynchronous, back-end access to the file systems upon which, or databases within which, the data resides. This is similar to the process through which data is backed up for system recovery purposes.

The type of digital material being archived and the ways in which it is currently created, stored, and accessed will tend to determine what approach or combination of approaches is appropriate to a particular archival procedure.

As an illustration of this methodology, we recommend that e-mail, for which the archival intent is selective rather than comprehensive, be captured by two different “push” methods:

- The first process would require the recipient of a message to forward that message to a special archival address or to copy that archival address on an outgoing message.

- The second process would enable archiving be automated into the workflow of message processing according to rules established for determining archival eligibility requirements. For example, the University could establish a policy of archiving all mail to and from officers of the University. This sort of rule could be implemented within the normal processing of mail.

In both processes, messages are actively being “pushed” to an archive.

On the other hand, the archiving of web sites, for which a comprehensive overview is desired, necessitates methods that take a “pull” approach.

- In one recommended process, a periodic “snapshot” of the University’s main web sites would be created using storage management technologies already in place. This snapshot will effectively create a replica of the web servers’ systems. Specialized software known as a “web crawler” will then traverse this replica as if it were being accessed by a user with a web browser. In doing so, this software will create a copy of the web sites that preserves the users’ view of the University web. This copy will then be stored in an archive.

- In a second recommended process, an archive of the web servers’ data would be made through a “point-in-time” capture of the replicated files and databases that constitute the web servers. This will preserve the web from the point of view if the server rather than the user.

These are both “pull” methods because they take data from the target system, in this case the web server, without any active participation on the part of the system being archived.

3.3.4 Media Preservation and Technological Change

There are two technical issues that face us regarding the media upon which the digital archive data is stored. The first is that media can fail and data can be lost. The second is that the media supported by the information technology industry changes as technology advances. Strategies can be implemented to address both of these issues.

Safeguarding Against Media Failure

We can say without fear of contradiction that no media is completely safe from data loss. Environmental conditions, mechanical failure, human handling, stray radiation, and any number of causes can lead to loss of data. Even under ideal conditions digital media will deteriorate over time leading to loss of data. Magnetic domains will weaken and even optical media will suffer eventual failure of the recording substrate. Fortunately, there are mitigating factors that mean that data need not be irretrievably lost.

The first factor is that the media and media recording and retrieval technologies are capable of recovering data that is not directly readable through reconstructive algorithms. These algorithms are commonly used and are continually being advanced through further research.

The second factor is that the IT industry has developed best practices for managing media to maintain its viability. Some of these best practices follow.

Media Management Best Practices

While soft, or recoverable, errors are expected and can often be dealt with automatically by hardware or software, unrecoverable errors will still occur. For example, a tape drive mechanism could develop a problem which physically damages a tape. Because of this possibility, the first recommendation that storage professionals make for safeguarding irreproducible data against loss is to make multiple copies of the data and then geographically separate the data to guard against disaster such as fire.

Another recommendation is to store digital media in a properly controlled environment and to carefully monitor the environment against failure. Temperature and humidity will significantly affect the useful life of any digital media, whether it is magnetic (such as tape) or optical (such as DVD and CD). Where a tape cartridge may be expected to have a useful life of ten years at a temperature of 20c and a 40% relative humidity, that life may drop to three years at 25c and 50% humidity.

A further recommendation is to monitor the occurrence of recoverable errors on each individual medium, to project the occurrence of recoverable errors, and to copy the data from media that is showing signs of wear to fresh media if the trends warrant such.

Finally, it is recommended that data periodically be copied onto fresh media. Monitoring for the occurrence of recoverable errors assumes that the media is regularly accessed. Much archival data may not be accessed for very long periods, if ever, once it is written to the archive. Therefore, an active preventative maintenance program is required to maintain media viability.

Safeguarding Against Changes in Technology

Even if all of the above-mentioned best practices are followed, the technology used to record the data will eventually become obsolete.

Hardware vendors understand that migrating data from one archival technology to another is a difficult and expensive process; therefore, the technologies used for digital archiving tend to be supported longer than other information technologies. However, no vendor will commit itself to support a particular hardware platform indefinitely, since the cost of maintaining old technology grows over time until it is no longer cost-effective for the end customer or profitable for the vendor. Support staff trained on the old hardware become hard to find, parts are no longer manufactured and, eventually, fresh media may no longer be obtainable.

Additionally, even though vendors may make an effort to support aging hardware, this generally cannot be said of software. Support for older software platforms upon which one may be reliant for management of one's archive is very likely to be withdrawn well before support for the hardware on which it is running.

Organizations needing to preserve data for more than ten years should plan on migrating their archives to a newer technological base every ten to fifteen years. The costs of the new hardware and software are likely to be the least significant expense in such an effort. The staff resources required and lost opportunity costs may well outweigh the capital investment. There is also the possibility of data loss during such a technology change, which some estimates place as high as 5% for an average migration project.

Fortunately, these problems may be somewhat mitigated if the best practice of periodically refreshing the current media is followed. In this case, the processes for moving the data and assuring that it arrives intact are already in place. Furthermore, the staff effort and time requirements should be feasible to estimate based on the experience of copying the data within the current technology.

3.3.5 Archival Formats

Original "native" format vs. an archival format

As the owners of web pages and other types of non-textual data progress to newer technologies and more dynamic formatting practices, consideration must be given toward methods of saving content to insure that it is available for future use.

All items stored in a computer have a logical format, or way that the computer understands the information. These are typically related to the tool that was used to create the information. One of the most basic formats used in the web environment is HTML, or Hyper Text Markup Language. Most static text on web pages today utilize some form of HTML.

Other methods used to display information on web pages include technologies like asp, java, cgi, and other programming-related formats. Objects utilizing these formats typically serve some type of dynamic function.

As we consider how to save or archive objects and text from web pages, we must consider whether to save them in their *native format* (i.e. java, html, cgi) or to convert them to some type of more *generic format* (basic text) which would be more accessible.

Formats will evolve over time, as will the use of those formats. Ten years ago, most web pages utilized static HTML to display their content. Today, more web pages utilize a combination of static HTML and dynamic objects. With the passage of time, it is highly probable that the percentage of static pages will continue to drop to a lower level.
The same can also be said for other technologies and formats. For example, images stored

in existing standards like jpeg have already seen movement to newer standards, such as jpeg2000. Video continues to evolve as compression and quality are increased. Research data may contain hardware-specific telemetry information. As the type of equipment is renewed or retired, that data may not be able to be deciphered.

3.3.6 Metadata

Metadata supports the basic archival functions of data ingest (deposit), preservation, discovery and dissemination (access). Metadata also supports records management retention policies. A balance must be found between not providing enough metadata, which would compromise the execution of basic functions and policies, and providing more than enough, which would be costly to provide.

This section begins the task of considering how to apply the relevant metadata standards developed by the digital library and learning object management communities to the five types of materials described in the charge so as to strike a balance between practical necessity and economic feasibility.

Introduction

This discussion is divided into three parts. The first describes the metadata standards for digital objects which librarians have developed during the past decade, some of which they are still developing. The second considers which of these standards might be relevant to the five types of materials described in the charge, because the types of materials on which librarians have so far focused are often of a different type (for example, so-called cultural materials). Therefore, some analysis is necessary before trying to apply standards originating in one domain to materials originating in another. Finally, we present a framework for establishing the core metadata elements needed for archiving University materials. These last two sections reflect the understanding that while some metadata are essential, there is a cost to metadata creation or capture which must be contained.

Metadata Standards

"Metadata is data about data. A good example is a library catalog card, which contains data about the nature and location of a book: it is data about the data in the book referred to by the card. The content combined with its metadata is often called a content package." --Wikipedia (http://en.wikipedia.org/wiki/Metadata)

Metadata has progressed a good deal from the days of the library catalog card. Those working with digital libraries today commonly recognize the following kinds of metadata:

- descriptive
- preservation
- rights

- technical
- structural

Descriptive metadata most closely resemble the bibliographic data found on a library catalog card.

Preservation, rights, and technical metadata (defined below) are often grouped together under the overarching term "administrative metadata." These can also include other kinds of metadata, such as digital provenance, which describes the history of a digital document, including its migration to new formats.

Structural metadata describe how a set of digital objects should be combined to form a compound digital object, For example, such metadata would describe how individual page images should be combined to form a digital book (if the book were scanned page by page), or how audio tracks should be combined to make a recording. The above definition of metadata added: "The content combined with its metadata is often called a content package." In practice, standards for structural metadata also include content packaging information as well.

Descriptive Metadata Standards

MARC (MAchine-Readable Cataloging)

The traditional standard for representing machine-readable bibliographic data is MARC, which describes both an exchange format (a syntax) and a markup specification (a semantics). The modern digital library replaces the MARC syntax with XML (Extensible Markup Language), and has introduced new descriptive metadata standards for digital materials. A brief introduction to some of the more important of these follows.

Dublin Core (DC)

Dublin Core exists in two forms: unqualified or simple (15 core elements), and qualified. Unqualified Dublin Core is required by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), an important new standard for facilitating the exchange of descriptive metadata. It is regarded as the least common denominator, or *lingua franca*, of metadata exchange, and it is important for that reason.

MODS (Metadata Object Description Schema)

MODS, like Dublin Core, was designed for the description of electronic information. It is richer than Dublin Core, but considerably less complex than MARC. It is very new, but has gained "mindshare" very quickly. It is likely to play an increasingly important role as a descriptive metadata standard for electronic information.

IEEE Standard for Learning Object Metadata (LOM), no. 1484.12.1-2002

This standard, now published and available for purchase on the IEEE (Institute of Electrical & Electronics Engineers) website, is also available online as a final draft standard document. LOM was developed to describe learning objects. It contains a mapping of its own elements to unqualified Dublin Core in order to facilitate basic interoperability with the larger digital library community.

Customized metadata

For some projects the Library finds it necessary to create customized descriptive metadata element sets. However, in these cases it, too, creates mappings between these custom elements to unqualified Dublin Core, to facilitate the exchange of metadata, for example, via OAI-PMH.

The Library's Non-MARC Metadata Working Group maintains information on other descriptive metadata standards and proposals which are also of importance or interest to the digital library community. Of these, one is of especial interest to us. One of the creators of Dublin Core, John Kunze, simplified that standard even further, as follows:

- who (in Dublin Core terms, "creator," more traditionally, "author")
- what (title)
- when (date of creation)
- where (a persistent or permanent identifier or locator, such as a persistent URL)

Though it is not a standard, but rather a methodically articulated proposal, this Electronic Resource Citation format, or ERC, which Kunze describes in more detail in *A Metadata Kernel for Electronic Permanence*, is worthy of our attention as presenting a cost-effective, core descriptive metadata element set which may prove serviceable enough for archival description.

Preservation Metadata Standards

Preservation metadata record information required for the preservation of digital objects. Core preservation metadata record information not recorded by another applicable standard (i.e., descriptive, rights, structural, or technical). A core preservation metadata set is currently being defined by the PREMIS (PREservation Metadata: Implementation Strategies) working group, sponsored by the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG). The final draft standard is expected at the end of 2004.

Rights Metadata Standards

PREMIS is also considering what kinds of rights information might need to be included in a core preservation metadata element set. The digital library community is several years away (at least) from a rights expression language suitable for its purposes. ODRL (Open Digital Rights Language) and XrML (eXtensible rights Markup Language) are too narrowly focused on digital media and commercial publishing interests. In the absence of any applicable standard, simple local rights expression languages may be developed to address local needs.

Technical Metadata Standards

Technical metadata answer the question: What kind of digital object is this? Possible answers might be, TIFF (a standard for digital images), ASCII (a standard for digital text), etc. Technical metadata should also give more precise information about the kinds of formats a digital object contains. For example, there is more than one kind of TIFF, PDF, etc. PREMIS is defining a core set of technical metadata for preservation purposes. The Library is also working to define its technical metadata requirements, which should be available soon. It will be compared to the PREMIS core set when that is available.

Because identifying and recording technical metadata can be expensive, automatic extraction is attractive as a way to keep costs in hand. JHOVE (JSTOR/Harvard Object Validation Environment), a tool for the automatic extraction of technical metadata from digital objects, is being developed to address this need.

Structural Metadata Standards

Three standards for structural metadata currently have "mindshare" in the Library community.

METS (Metadata Encoding and Transmission Standard)
"The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation." --
http://www.loc.gov/standards/mets/ METS is the structural metadata standard with the widest acceptance in the digital library community today.

MPEG-21

MPEG-21 is an ISO standard which can serve as an alternative to METS. It is being used by one digital library, the Los Alamos National Laboratory Research Library, which has pioneered the use of OAI-PMH and other standards, and which plays an intellectual leadership role in the digital library community. The difference between METS and MPEG-21 may be summarized as follows: METS is tailored to the needs of the digital library community today; MPEG-21, if it sees wide industry adoption, may result in a wide array of tools tomorrow which would then be uneconomical for the digital library community to ignore.

IMS Content Packaging Specification

"The IMS Content Packaging Specification provides the functionality to describe and package learning materials, such as an individual course or a collection of courses, into interoperable, distributable packages. Content Packaging addresses the description, structure, and location of online learning materials and the definition of some particular content types." -- http://www.imsglobal.org/ There is also an IMS Meta-data Best Practice Guide for IEEE 1484.12.1-2002 Standard for Learning Object Metadata (the IEEE LOM, described above).

Metadata Standards Applicable to Archiving University Materials

The charge asks us to consider five types of materials:

- University Web
- E-mail
- Instructional Materials
- Administrative Records
- Research Datasets

We will apply the five types of metadata to these five types of materials, to see which are applicable.

Archives whose purpose is the preservation of digital materials differ with respect to how much descriptive metadata to record about deposited materials (hereafter, "deposits"). A minimalist approach is taken by the oldest of these digital archives, that of Harvard University. It asserts that full descriptive metadata should reside elsewhere (for example, in the university's online catalog). The archive keeps only minimal descriptive metadata, sufficient to help match up deposits to a canonical description in an external catalog. This helps answer the question, from the archive's perspective, "What object is this?".

We agree with this approach. We assert that good bibliographic description is important, but we also assert that what constitutes good bibliographic description may go beyond what an archive needs. A reasonable, core descriptive metadata element set is provided by Kunze's Electronic Resource Citation (ERC), which maps easily to unqualified Dublin Core for interoperability (e.g., the export of metadata records into another system), but which at the same time identifies a "core Dublin Core" for the purposes of archiving University materials. In looking at the five types of materials, it would seem that ERC's "who," "what," "when," and "where," or (in Dublin Core terms), "creator," "title," "date," and "identifier," apply to all of them. However, having defined, or required, core descriptive metadata elements for deposits does not thereby disallow fuller descriptive metadata elements from being provided if they exist; for example, one can easily imagine future deposits of instructional materials being already provided with full LOM descriptions. What the archive can do in these cases is extract the core elements it needs for its purposes (for example, using the LOM to Dublin Core mappings which the LOM standard defines), and store the full descriptive metadata together with the deposit as part of a content package.

The PREMIS core preservation metadata element set is currently under construction and due to be released as a final draft by the end of 2004. The group's membership is international in scope; it is being sponsored by both of the big bibliographic utilities (RLG and OCLC) in the United States, and it has reviewed all earlier work done in this area with a view to bringing it to conclusion. We must therefore look to PREMIS when defining our core preservation metadata elements, because PREMIS will set the standard in this area.

Because no applicable rights metadata element sets exist, we will have to construct one. We discuss this in the next part. We regard a definition of rights as fundamental to archiving all five types of materials.

The PREMIS group is defining core technical metadata in addition to core preservation metadata. This work needs to be tracked, because technical metadata are crucial for the survival of deposits over the long term. Our position is that, given the considerable cost of manually inputting technical metadata and the insurmountable cost of requiring that these be provided up front for all deposits, the archive should rely on automatic extraction of technical metadata for its deposits, using tools such as JHOVE.

With the exception of instructional materials, we do not think that structural metadata standards are important for the types of materials identified in the charge at this time, though they are important for many traditional digital library materials. In addition, they might become important in the future, as digital archives mature and as one of these standards is required for the dissemination of content packages (data and metadata); they are certainly unavoidable today for archives whose purpose it is to serve the digital library community. For instructional materials, IMS may become more immediately important. Structural metadata might become more immediately important if archiving University materials becomes so successful so quickly that a demand is created for this activity to expand beyond its original scope. For example, there could be pressure for an archive to serve the purposes of an institutional repository to facilitate scholarly communication (something which today is served by systems such as D-Space from MIT, or GNU EPrints from the University of Southampton). Though simple documents, such as PDF files, which constitute the bulk of what institutional repositories contain today, do not need content packaging for their dissemination, compound or complex documents do.

A Framework for Establishing the Core Metadata Elements Needed for Archiving University Materials

The following is informed (though not exclusively) by discussions in the PREMIS group, and also by discussions in a subgroup of the Library's Archiving Group charged to specify the functional requirements for a Library archive, which are also informed by PREMIS. Because PREMIS itself is informed by prior work, and because the following is a synthesis which includes our own thinking, we do not credit the origin of any idea in this part.

The archive is viewed as involving the following entities:

- events (or actions)
- agents
- objects
- rights
- relationships

Agents and objects participate in events. Events occur at definite times. Events are enabled by system functions or toolsets, and are governed by policies. Events may be likened to verbs; we identify two event modalities (or adverbs), "may" and "must" (i.e., optional and required). Policies determine what events may, must or may not occur. An archive should record both policies, to determine what events may, must or must not occur, and occurrences (e.g., this event involved this agent and this object at this time).

Events

The four core events involving agents outside the archive are:

- deposit (ingestion of materials into the archive)
- deletion (withdrawal of deposits from the archive)
- discovery (access to metadata about deposits)
- dissemination (access to the deposits themselves)

Upon deposit, the following events may occur:

- initial fixity benchmark (e.g., an MD5 or SHA-1 message digest, or fingerprint)
- virus check
- normalization (conversion of an object from one format to another which has more assurance of long-term preservation)
- compression (using a lossless compression scheme, such as gzip or bzip2)

After deposit, the following events either may or must periodically occur, as indicated:

- compression (if this did not occur on ingest: optional)
- duplication (or transfer to another physical medium: required)
- fixity check (determining whether the bits have changed: required)
- integrity check (for compound or complex digital objects, determining whether all the parts are present: required)
- migration to new formats (mandatory where the long-term preservation of an object is threatened; optional otherwise)

Events may occur as follows, as determined by policy:

- scheduled, i.e., according to a pre-specified periodicity
- always (i.e., anytime, or ad hoc)
- never
- before [date]
- after [date]

Agents

Agents should be recorded as structured elements consisting of at least these components: name (forename; surname); title (i.e., something that designates a function within an organizational unit); organizational unit (or affiliation). At the time of deposit, these values should correspond to (and be able to be validated against) a University directory, which is also archived with sufficient periodicity to allow tracking changes to these values for any agent and to allow future events to take place as expected.

From the perspective of the archive, the following types of agent exist:

- owner (of an object, or who it is that claims intellectual property rights to it)
- depositor (of an object)
- archive administrator, or the archive itself (whose domain is all objects in the archive)
- member of a group (or affiliations of agents)
- member of the world (i.e., an agent for whom identification, authorization and authentication is not required)

Groups (affiliations) may include faculty, instructor, researcher, staff, student, organizational unit, etc. An agent's type determines its rights to participate in specified archival events involving specified archival objects.

Objects

There are three kinds of object in the archive: data, metadata, and meta-metadata. Data are the objects themselves. Metadata describe properties of objects; they may include records of events involving those objects in the archive. Meta-metadata define who created the metadata describing the objects, when the metadata were created, and whose intellectual property they are. Meta-metadata are not only important for answering questions about metadata, but also to prevent the unauthorized use of metadata should an archive begin exchanging metadata with others, for example, using OAI-PMH.

Data objects participate in events. They need persistent (i.e., system-independent) identifiers to allow them to be unambiguously referred to by descriptive metadata. This is especially important if the primary finding aid (i.e., the catalog) for digital objects is not necessarily the archive itself, as we are recommending. Several well-recognized schemes for persistent identification exist, e.g., CNRI's The Handle System. However, locally created unique identifiers are also acceptable using simple mechanisms such as HTTP server-side redirects, which one report suggests scale well.

Rights

Rights policies specify:

- which agents may participate in which events
- which objects may participate in which events
- when an event may or must occur, or when it may not

To the extent that policies exist and are recorded programatically by the archive, for example, as actionable metadata, archival events can occur automatically. Otherwise, intervention by an archive administrator is required to determine whether an event may, must or must not occur.

The archive should be able to refer to a record of the reasons behind policies, to allow questions to be answered about them, and to inform any policy-review process.

Relationships

Some relationships are implicitly recorded by an archive. For example, a record that an agent and an object participated in an event implies a relationship between the agent and the object. However, some relationships need to be explicitly recorded.

For example, if, by policy, successive editions of a document (such as the University statutes) are to be archived, then it is important to record which is the latest version. Alternatively, if an identical document exists in two formats, e.g., PDF and ASCII text, then it is important to record that fact. This implies the need for a relationship element in metadata.

Turning this framework into a core metadata element set is a next step in the process of archiving University materials.

4. Specific Issues from Use Cases

4.1.1 E-Mail: Summary of Issues

E-mail has a number of characteristics that may aid or hinder the task of archiving it.

Beneficial Characteristics

Header Information

Each e-mail message contains header information. This header information is prepended to the message contents and has multiple purposes. First, it is used by the mail delivery agent to deliver the message. Second, it contains information that is useful to the recipient to identify the message. Third, it is used by e-mail client programs to sort, weigh, filter, or otherwise process the mail messages. Some of the information that can be found in the header of an e-mail message includes:

- The sender of the message -- both name and e-mail address.
- The recipient address of the message.
- The date of the message.
- The subject of the message.
- The mail servers involved in the delivery of the message along with timestamps.
- Threading information which records the mail messages that were involved in the message exchange.
- The status of the message (if it is retrieved directly from the mailbox) which will signal whether the message was read or replied to.

Distinct Ownership and Permissions

By design, e-mail belongs to a recipient and permissions are simple -- only the recipient has access to the contents of the mailbox. Therefore, ownership and access are straightforward. However, as is noted elsewhere, there are problems in determining changing roles (i.e., mail sent to Geoffrey Stone as Provost could be considered the property of the office of the Provost rather than the recipient personally) and the fact that roles and entitlements within the university are fluid as people arrive, leave, and change positions.

Detrimental Characteristics

High Noise Content

For the general recipient, a large percentage of the incoming mail will not be suitable for archiving. There are a number reasons for this, including:

- Delivery of "spam" or junk mail. Recent estimates, including our own tests with an anti-spam product, place spam as a percentage of the total volume of mail received at approximately 65%. Therefore, if unfiltered, over half of the mail that is delivered has no value to the university as archival material. Before the end of Fall quarter, 2004, there will be anti-spam filtering in place for the campus. However, even with such filtering, there is likely to be a significant percentage of spam in local mailboxes.

- Delivery of personal mail. Most users receive personal mail at their University of Chicago address. This is accepted by policy and should continue to be. However, such messages are not suitable as archival material for the university and, for the sake of individual privacy, all reasonable precautions should be taken to keep such material out of the archive.

- Delivery of "irrelevant" mail. Many messages that are sent to our users may involve official business of the university, but may not be important in any sense for archival purposes.

As a consequence of the large number of messages unsuitable for archiving, the size of the mail archive itself can be quite large. This situation will have an impact on both the cost of storage and the difficulty in identifying and retrieving useful information from the archive. Therefore, with regard to mail, certain remedies should be implemented to minimize the number of irrelevant messages in the archive. Two remedies that should be considered are:

- Anti-spam and anti-virus filtering. NSIT is currently in negotiations with a vendor for a product that will address these needs.
- Requiring human decisions to enter mail into the archive.

Lost Messages

Mail messages are lost between backups. The same will be true of archival runs. Mail messages are lost because:

- Mail messages may be received and deleted, deliberately or inadvertently, between successive archival runs.
- Mail messages may be downloaded to the message recipient's desktop and not archived as a result.

Possible remedies for this problem include:

- Cloning received messages to an archival site.
- Frequent "snapshot" backups of data that are retained until processed for missing mail messages.

Special Data Formats

As e-mail has increased in reliability and popularity, and as computer technology has become more versatile, powerful, and affordable, e-mail has become a medium for transmitting a wide variety of file and attachment types. Messages may have attachments that include:

- Office application files (Word processing, spreadsheet, presentation files)
- Multi-media files (Audio and/or video files)
- Research data (Statistics, raw data, images)

In order for the archival material be useful to future users, the archive must take into account the fact that the message must, at a minimum, be split into its constituent parts. Therefore, the archivist must have access to software that will continue to split any given message.

Once the message is split, the archivist must be able to utilize or process the constituent parts for them to be useful. This can be problematic for a number of reasons, including:

- Computer technology advances rapidly and file formats are developed, altered, and discarded over time. For instance:
    - Microsoft Word documents have continued to evolve over time. Older versions of the program will not open files generated by newer versions of the program.
    - GIF image file format that once was ubiquitous has been displaced by JPEG as an alternate format. It will probably fade away as a result of licensing on the code that reads this format.
    - WordStar was once an extremely popular word processing software which is not currently used much.

- There is a rapidly increasing number of programs and formats that may be used to generate the files that are attached to an e-mail message.

There are a few possible remedies that can be implemented to address these problems:

- Tools can be added to the archive service to read common file types and those tools can be reviewed, enhanced, and revised annually.

- Formats that are accepted into the archive will be limited to specific known standards for which tools can be made available. This remedy can be imposed upon the user by requiring them to convert the file before archiving the message or can be part of the process of archiving itself. As a message is processed into the archive, tools can convert the message attachment to a supported file format and store it along with a copy of the original format.

- For uncommon file formats such as research data, metadata describing the file format and possibly containing uncompiled program code to read the file can be required at the time that the message is submitted to the archive.

Unfiled Structure

E-mail messages, unless some intervention occurs, are delivered to a common mailbox. As such, they are not filed into folders that would group them according to subject. Therefore, unrelated messages will be intermixed, making it more difficult for the archivist to discern relationships between messages.

Some remedies that could address this problem include:

- Have the recipient sort the messages at the time of submission.
- Allowing mail message "threading", which is a common utility for most mail programs, to help identify message relationships.
- Require a submission form fully identifying the context of the message at the time of submission.

4.1.2 E-Mail: Current State

E-mail to the University of Chicago network is delivered to members of the campus community through a number of paths. These paths can be divided into two categories:

- E-mail that is delivered to and sent from the NSIT mail server.
- E-mail that is delivered to and sent from other e-mail servers.

NSIT Mail Service

Architecture

The NSIT mail service consists of an integrated group of approximately a dozen machines that handle the various aspects of e-mail delivery. These functions include:

- Mail relaying. Servers performing this function accept mail from other computers that are not equipped to handle mail delivery. An example of computers requiring this service would be a desktop computer running the Microsoft Windows operating system and Eudora.
- Mail serving. Servers performing this function store mailboxes and allow e-mail clients (for example, Eudora) to access and alter those mailboxes.
- Mail forwarding. Servers performing this function accept mail and forward it to other mail servers.
- Mail list serving. Servers performing this function accept mail and forward accepted mail to all members of the recipient list. A mail list may set to save a copy of all mail received which can be accessed through a web browser.

- Webmail service. Servers performing this function access mailboxes stored on a mail server be means of web-based programs that can be accessed through a web browser.

The systems that constitute this architecture are manufactured by Sun Microsystems and run Sun's Solaris operating system.

The applications software that perform the functions outlined above are all open source software -- meaning that the source code is available to the public to alter for the environment in which it is installed. The applications software is installed, configured, and maintained by NSIT systems staff. It has, in some cases, been altered to accommodate university policies or procedures.

This system is about to undergo a radical restructuring over the next couple of months. It will be replaced with a commercial product that will run on an Intel-based hardware. It will include all of the services listed above and will also include spam and antivirus filtering capabilities. Though the services provided by this new system will be in many ways enhanced, NSIT will not be as able to customize this new system to address local policies or practices.

Mail Flow

NSIT's mail service currently handles approximately 63% of the outgoing mail connections and approximately 53% of the incoming mail connections that transit the university's network. It must be noted that the 53% represents mail connections into the university from outside the uchicago.edu domain and does not include intra-domain mail. Also, it includes any connection to the “midway” domain that has a *user*@uchicago.edu address and the final destination may not be the NSIT mail server.

For most mail traversing the NSIT mail service, the entry point for incoming mail is a set of systems collectively known as midway. All mail addressed to *user*@uchicago.edu or *user*@midway.uchicago.edu is accepted by midway for processing. In the case of mail addressed to *user*@midway.uchicago.edu, the mail is forwarded to the mail server (plaisance). Plaisance stores the mail into *user's* mailbox on its attached disk storage units unless mail forwarding for *user* is enabled. If mail forwarding is enabled, the mail message is forwarded to another mail server.

For mail addressed to *user*@uchicago.edu, midway queries the campus directory server to resolve an address for *user* and forwards the mail accordingly. The mail may be forwarded to any mail server on the network. Approximately 75% of @uchicago.edu addresses forward to plaisance.

Mail is delivered to each mail client by setting the client application to use either the IMAP or POP protocol. The client can query plaisance and interact with the mailboxes stored there. Using the client *user* can:

- leave mail messages on the server

- delete them
- forward them to others
- download a copy to their local computer
- transfer them to their local computer
- make copies of messages on their local computer
- file them in folders on the server

Mail sent from campus desktop computers configured with NSIT's recommended settings or from desktop computers which have the NSIT Connectivity Package installed are configured to use NSIT's mail relay servers to forward e-mail messages for delivery. All mail sent from such computers will traverse these systems.

Mail clients may be configured to file a copy of any mail that is sent to a special file. This file can be on the local system or may be a mailbox that is stored on plaisance.

Current Backup Situation

NSIT's e-mail service currently is backed up to magnetic tape by a dedicated robot. All data on all of the constituent systems are backed up each night. The tapes are erased and reused approximately every two weeks. There is no long term archive of any mail that is stored on the system.

The backup tapes contain the mailbox files as they were at the time of the backup. The following information is stored on the backup and can be used to generate metadata that would be useful for true archiving:

- From the mailbox
    - The recipient of the message
    - The sender of the message
    - The recipients on the Cc: line
    - The date of the message
    - The subject of the message
    - The message text
    - Status flags indicating whether a message was read and whether a reply was sent

- From the file system metadata
    - Access permissions on the file
    - File modification, access, and creation dates

It must be noted that the information recorded as part of a backup will lose its context. For example, a person's role may change, the membership of a mailing list will not be recorded with the content of a message, etc.

Campus Distributed E-Mail Services

Architecture

The e-mail architecture for the communities and individuals who do not use the NSIT mail services is heterogeneous and often unknown. The architecture consists of:

- A few departmental servers that support 50 or more users. (Examples include BSD/IS, and the GSB.)
- Numerous servers that support small sub-departments.
- An unknown number of desktops that are capable of accepting, delivering, and storing e-mail and are used to doing so.

Mail Flow

Mail flow to non-NSIT mail servers is unknown. The GSB is the next largest mail service on campus, receiving about 18% of the incoming connections. The third largest receives 3%.

Current Archival Situation

The current archival situation is unknown. The larger departmental servers undoubtedly make backups of their systems, but retention policies vary. Smaller systems and desktops may or may not backup data at all. There is a possibility that NSIT's TSM backup system may be used to backup data.

4.1.3 E-Mail: Concerns

The Archival Flag

If an archival flag is to be set, that flag leverages the infrastructure that is already in place. The recommendation is that it reside in LDAP (Lightweight Directory Access Protocol). The reasons for this are as follows:

- LDAP is already installed and used by NSIT's current and future e-mail service infrastructures.
- LDAP is already integrated into NSIT's account claiming and similar software. We already have web-based programs that allow the user to easily edit the content of their directory information, and these programs could be easily adjusted to permit the setting or unsetting of this flag.
- LDAP is extensible, so adding this flag is not difficult.

E-mail Server Processing Logic

The replacement e-mail server infrastructure that NSIT intends to have in place before Winter quarter, 2004 will permit a copy of an e-mail message to be forwarded to an archival e-mail system based upon values set in the LDAP directory, or upon patterns that are matched in the message itself. The advantage of using a pattern is that it can lend flexibility to the archival process. Rather than all messages or no messages being archived, some messages can automatically be archived. Patterns can be matched anywhere in the message. The most likely locations for a pattern in the message are:

- The domain name of the sender, recipient, or reply-to. (*user*@archive.uchicago.edu)
    - Pros:
        - It can be made part of a "personality".
        - Domain names as part of an e-mail address are pretty well understood by users, making manual use of them feasible.
    - Cons:
- A filtering pattern added to the recipient portion of the sender, recipient, or reply-to. (*user*+archive@uchicago.edu)
    - Pros:
        - It can be made part of a "personality".
    - Cons:
        - The "+" syntax is not generally known to e-mail users and so it is difficult to instruct users about manual use.
- A tag on the subject line.
    - Pros:
        - Easily added by the user.
        - Persistent. Any reply will have the pattern placed on the subject line as well.
    - Cons:
        - Intrusive on the subject line.
- A special header line.
    - Pros:
        - Not generally visible to the recipient, so not distracting.
    - Cons:
        - Only works on outgoing mail.
        - Can be difficult to teach a user to set this.

Possible Options for the Archival Server

There are several options that can be considered for the archival service:

- NSIT can custom-design a solution with NSIT personnel or contract programmers.
    - Pros:
        - Can be customized to our environment and needs.
    - Cons:
        - Extremely expensive.
        - Diverts resources from other projects.
        - Takes time to develop.

- NSIT can add extra servers of the type being purchased to replace the current e-mail service infrastructure.
    - Pros:
        - Can be managed by current staff who will be managing the e-mail system.
        - Relatively inexpensive.
        - Commercial support.
    - Cons:
        - Requires development of customized filters.
        - Not really an archival solution, so it does not address a number of needs.
        - Not a generalized solution.
- NSIT can purchase an archival e-mail-only solution.
    - Pros:
        - Does not divert personnel.
        - Addresses archival needs.
        - Commercial support.
    - Cons:
        - Not a generalized solution.
- NSIT can purchase a general archival solution and use it for e-mail archiving as well.
    - Pros:
        - Does not divert personnel.
        - Addresses archival needs.
        - Commercial support.
        - Generalized solution.
    - Cons: $4140.00

4.2.1 Web: Summary of Issues

The world wide web is a hierarchical and network-available protocol for the widespread distribution of multimedia content. It is easily accessible and is becoming one of the primary methods of communication.

Beneficial Characteristics

Hierarchical Construct

The web, by design, is hierarchical. Therefore, information and documents generally are filed into groups of related content. This can facilitate searching an archive to find information.

Detrimental Characteristics

Links May Point to Offsite Resources

The ability to link to content on remote servers makes the web a flexible and powerful agent for the dissemination of information. However, this same quality could lead to problems for archival processes. First, there is a policy and legal question about archiving documents from another site. Secondly, there is the fact that the links can cascade or chain to other sites, making the amount of data that is archived very large.

Access Lists Can Be Problematic

Access lists, which grant access to web-based documentation to users or groups of users are a problem for the following reasons:

- The archived access list remains static in time, while the population of users that should have access is fluid.

- The functionality of the access lists, which are stored as files within the web-hierarchy or in the configuration files for the web server application, are tied to authentication methods which are tied to services that are not part of the web structure that is being archived. For example, access to student records requires a person to authenticate by using a network identification token (CNetID) and a password. The authentication information is not stored on the server, but rather within an authentication server. In the case of some applications, further authentication information may be required. For example, a person's role may be queried from the campus directory server to determine whether the authenticated individual is categorized as being a "student" in order to be granted access to a portion of the site.

- Web spiders, a methodology that can be used to “snapshot” a website, capturing content by traversing it, will only be able to access content that is available to the general public. It is possible to set access rights on the web site to allow the computer running the spider to have access to the entire site and all content, but by doing so, the access information is lost. The spider will collect all data on the site, but will have no stored information about the access control.

Formats Used By Web Are Not Static

As with e-mail, there are many data formats that are used by the web. Those formats are diverse and non-static. Once archived, it is important that the data remain usable. Therefore, part of the design of the archival infrastructure must include methodology for making use of archived data.

Amount of Data Archived Can Be Substantial

The interlinked nature of the web along with the types of data that are distributed through it will ensure that the size of the archive will grow at a considerable rate. Our estimates place this growth rate at about 40% per year. The growth rate is tied to several factors including:

- The growth of the web as a marketing tool. As the number of connections to the Internet have grown and end-user familiarity with the web has increased, organizations recognize that their product can be marketed by the web.
- The low cost of distributing information via the web versus paper documentation. As computers and storage have become more powerful and less expensive, the web has become an inexpensive method for distributing documentation as compared to the traditional methods of paper documentation and bulk mailings.
- The growth of multimedia. As a result of the improvement of network capacity and digital production techniques, the web is increasingly used to distribute multimedia presentations. On demand audio and video are now a viable option to printed documentation

4.2.2 Web: Current State

NSIT Web Services

Unix-based Services

The largest portion of NSIT's web services, whether measured by number of sites or amount of data stored, resides on Unix systems. Currently, these systems are on hardware supplied by Sun Microsystems and run the Solaris operating system. The web server application is the Apache open source web server, which is the most commonly used web server application on the Internet. All data on these systems are backed up nightly to an Exabyte Mammoth 2 tape library that is shared with the e-mail system. Like the e-mail system, there are 6 incremental backups and 1 full backup done each week. A full backup copies all of the files on a system to the tape device, while an incremental backup copies only those files that have changed since the last full backup. The tapes on which the backups are saved are recycled about every 2 to 3 weeks.

Windows-based Services

Many highly critical services, including student registration services, reside on NSIT's Microsoft Windows-based servers. The servers run on hardware from Dell Computers running a version of the Microsoft Windows operating system. The web server application used is Microsoft's Internet Information Server (IIS). IBM's TSM software and IBM tape robots are used to backup the data stored on these servers.

Information General to Both Systems

Both the Unix-based and Windows-based systems archive data from the servers directly from the file system, thereby bypassing the web server. By doing so, the backup system retains all metadata relating to the files and the file system – including the dates of creation, modification, and access, the individual and group ownership of the files and directories; and the access permissions of the files. Therefore, the backup system is able to access all of the files on the system and to access it in raw format, unlike web crawling programs that access files through the web server and must accordingly deal with access permissions set within the server and receive only processed web pages.

Distributed Web Services

In addition to NSIT's services, many departments and divisions within the university operate web servers of their own. These systems share a number of characteristics that must be dealt with in order to incorporate them into a reliable and robust archival infrastructure. These include:

- The distributed environment itself is heterogeneous.
    - The servers run an assortment of operating systems on an assortment of hardware architectures.
    - The servers may utilize various web tools to present their pages. Examples include Java, Java Server Pages, PHP, Perl, Python, Active Server Pages, etc.
    - The servers are administered by various systems and web programming staff having differing levels of expertise and may not communicate with other divisions or departments on campus.
    - The web servers run a variety of web server applications.
- The content served by the web servers and the policies enforced by them are unknown.
    - Ownership of the files is unknown.
    - Authentication method is unknown.
- Whether a server has a process for backing up data resident on the system is unknown.

4.3.1 Administrative records: Summary of issues

The realm of administrative records for the University encompasses a wide variety of data types. This data can be found in the guise of paper forms, reports, and other assorted documentation. It also is found in the databases and files of the systems which support the University business offices. This data is stored in digital format, and is what is being addressed in this section.

4.3.2 Administrative Records: Current State

Digital Administrative records, as maintained in the University's systems, are today kept primarily on servers and large mainframe computers managed by NSIT. These computers are typically housed in a secure environment to protect the valuable information maintained on them. This data is usually accessed by graphical or textual interfaces which are written by University staff or are provided as part of the packaged software application by vendors. Reports are generated from these data stores, formatted by programs that either print them out on paper for distribution or provide access via a desktop computer for viewing in some manner.

Some of the data stores are maintained in record format in large files. Others are kept in databases, which store the data in a format utilizing a relational, or table structure.

Backups of these data stores are done to insure a reliable guarantee of the data in case of a catastrophic failure. These backups are usually done on a regular schedule, and are maintained in a proprietary format. Backups are not usually maintained indefinitely, but instead expire and are replaced based on schedules determined by the data owners. There is no type of descriptive data supplied with the backups. These backups are only accessed by technical people associated with the application owners who need to recover the data due to failure or error.

Until recently, the data stores contained data which could be thought of as being only text-based. However, with the recent advent of digital imaging of records such as time cards, purchasing documents, and other hand written information related to the business processes, graphical data is now maintained as part of the administrative record data store.

4.3.3 Administrative Records: Concerns

Archiving data from administrative systems will require a different mind-set than the current practice of simply backing up the data stores. In this regard, the following issues will obtain:

Legal requirements

In some cases, data may be protected by legal doctrine or precedent. Legislation such as FERPA (Family Educational Rights and Privacy Act), HIPAA (Health Insurance Portability and Accountability Act), and other personal data guarantee laws could affect who and how access to the data would be facilitated from an archive in the future.

Metadata

Metadata, data that facilitates knowledgeable access to the data stored in the archive, would need to be extracted, created, and formatted for each data category that is being stored. Generally, this metadata would fall into the following categories:

- schemas
- file layouts
- data definitions
- context definitions
- code books/valid values/lookup values

Extraction

Depending on the format of the data store, different techniques for extracting the data would be employed. The formats of the original data stores of the systems (i.e. file-based, relational database, etc.) are designed to provide maximum efficiency of the methods used to access and update them. This means that all of the data associated with a particular entity, transaction, or other data record will most likely not be stored in a contiguous manner. Scripting or programming may be required to extract the data in a way that maintains individual record or table referential integrity. This scripting will result in a format where all data is related to each other.

Archival format

Backups that are used today put the data into a proprietary format that is specific to the technology being employed to store the data. Each vendor's database and file format are written to the storage media in a way that guarantees that utilities used by technical personnel can successfully read and recover that data in an expeditious manner. These formats are not available to tools outside of the technical environment used to store the data, and therefore would not be able to be accessed by tools that a future archivist or other research personnel would use. This means that the data must be extracted in a format that is more generally available to search and viewing environments. The most common format would be text-based, allowing a multitude of tools to access the data.

Access

Future access to the archived data would be influenced by the following factors:

- Organizational structures and associated permissions to view the data would be an important issue. Some of the data in the archive could be of a sensitive nature (payroll, personnel, health, etc.) and thus either protected by personal privacy statute or policy. As is necessitated by business practice, certain data can be viewed by appropriate organizational personnel for business use. The correct safeguards would need to be in place to insure that the data is safeguarded.

- Proper selection of the appropriate tools to access the archived data would be important. Keeping in mind that the data, although in a standardized textual format, is still most likely archived in a structure that corresponds to a file structure or table schema, tools that would allow query of the archived data utilizing the natural structure would be required. This would allow associated records to be joined together with their related data records.

- Access to the data in the archive would be useless without a suitable level of knowledge of the data structure and content. An understanding of the relationships between various types of records, along with the required lookup tables and other codes used in the data would be necessary to make any valuable sense of the data content. In all cases, the required data dictionary, schema, file layouts, and other associated metadata would be essential to make use of the archive.

4.4.1 Course Material: Summary of issues

Chalk, the campus learning management system (LMS), represents the largest single identifiable repository of course content on campus. Despite this, there are a number of independent course sites that still exist, and several departmental and divisional servers that have no clear course-related function, but still serve an instructional role by hosting content that crosses several domains including research and teaching. In the broadest sense, the issues surrounding the archiving of course material include storage, but also raise questions of the reuse of, access to and ongoing maintenance of course content. There remains further the issue of Chalk as an academic administrative system that houses information on specific course-related student activity, participation, and contribution.

4.4.2 Course Material: Current state

Chalk: Campus Learning Management System

The Chalk Project began in late 1998 as an effort to identify a campus-wide LMS that could address two specific needs: to reduce the overhead associated with creating and maintaining a course website and second, ease the amount of administration associated with conducting a course. Selected by an ad hoc faculty group, the Blackboard Learning Management System and its successor, the Blackboard Academic Suite, has provided a stable enterprise Web-based framework for courses across all schools and divisions. It has become a major repository for campus course content, and, unlike many of our peers who struggle with managing several systems across multiple departments and schools, is the only LMS in widespread use on campus.

Today, Chalk represents a suite of systems that are housed within the NSIT machine room at the 1155 Building and co-managed by Academic Technologies and the Data Center. The entire environment operates on enterprise-class Sun hardware running Sun Solaris (Unix) and Oracle, and is made up of three distinct system components: learning system, portal, and content system. At the core of Chalk is the Blackboard Learning System which is the most familiar component of the three. On campus since AY1998/1999, the Learning System provides a unified course management infrastructure for faculty and a common online learning experience for students. The Blackboard Portal was added in Summer 2004 as a tool to simplify access to courses and to provide greater flexibility in tuning the Chalk user experience. The last piece, the Blackboard Content System, will be rolled out over the 2004/2005 academic year to provide simplified content publishing, content sharing across courses, digital rights management for course materials, and, with the Library, managed access to electronic reserves. The three systems combine to form a tightly-coupled suite of tools that provides the common online learning framework for campus.

In general, individuals think of a learning management system as merely a content repository, activity manager and delivery mechanism within a single environment. However, an LMS also serves an administrative function as it records and tracks student progress, provides a platform for electronic assessment, and acts as a venue for expressing ideas. As a result, Chalk can be considered a system of student record because it can and does track in some detail the work done by students within a specific course over a specified period of time. One must consider the administrative side as well as the academic side when assessing the archival needs of an LMS such as Chalk.

For the period between AY 2001/2002 through July 2004, a total of 2,933 courses were hosted on Chalk with an aggregate total of nearly 98.6 GB of course content. Statistics on the three complete academic years within the period are outlined below:

	Number of Hosted Courses	Total Content	Avg. Course Size	Avg. Growth in Course Content
AY 2001/2002	640	12.6 GB	19.8 MB	--
AY 2002/2003	1008	29.8 GB	29.6 MB	49%
AY 2003/2004	1189	51.7 GB	43.5 MB	47%

Chalk content and its Oracle databases are backed up daily for system recovery purposes in the event of hardware or software failure. Course content is initially archived at the end of each quarter to an off-site storage system located in the Regenstein Library and later transferred to DVD for long-term storage.

Hollywood: Campus Streaming Media Server

Although joint research into video streaming technology had been conducted by the University, Argonne, and IBM during the early-to-mid-1990s, a general-purpose streaming media server did not appear on campus until 1999. Recognizing a campus need but acknowledging that the technology was still not mature and interest widespread enough to justify a major campus investment, a small project was launched to address the slowly growing desire for a central video server and the result was Hollywood, the campus streaming media system.

Hollywood is a Sun Solaris (Unix) system that takes advantage of the Apple Darwin open source streaming server. Managed by the NSIT Data Center, Hollywood runs on enterprise-class Sun hardware and is regularly backed up, but the content is not archived. Access to the content on Hollywood is available to anyone, but publishing is limited to clients of the NSIT Digital Media Laboratory (DML). Materials are processed and compressed within the DML and transferred to an internal server prior to publishing. The client then submits a web form that collects appropriate basic metadata, invokes the publishing process whereby the content is copied from the internal server to Hollywood for streaming, and publishes a URL that can be copied into a Chalk site or a departmental web page.

Hollywood has been used in teaching geophysics courses and hosting grand rounds materials to delivering over 100 video streams to campus offices and departments of the inauguration of President Randel. Over the last five years, approximately 300 GB of low-quality streaming media has been hosted on the server.

Local Web Servers

Before the widespread adoption of Chalk, faculty, departments, and schools scrambled to publish course-related materials on the Web. As a result, a number of faculty and units hosted course materials on departmental or personal web servers managed by local IT staff, interested faculty, or students. Although Chalk has provided a platform for course content hosting and several units have migrated to it, a large number of these sites are believed to still exist as the legacy content is deemed too difficult to move, the servers provide unique capabilities beyond what Chalk can provide such as simulation and data collection, or the groups have chosen to maintain complete local control over the content and resources. Separating course content from other material is difficult as the local server may be hosting other materials, sites, and activities that may be related to departmental administrative needs, outreach programs, and/or research.

The technology, management, and backup policies and procedures (if any) vary widely and are unique to each server.

Faculty Websites

Aside from the issue of local Web servers, a number of faculty publish content within their personal sites on home.uchicago.edu. These materials are subject to the normal backup procedures of Home, but are typically not considered within the realm of course content. Like local web servers, it is difficult to disaggregate course content from other material as it may overlap with research, individual needs, and related programs.

Student Home Pages

As learning shifts toward being more group-oriented, the Web becomes a natural venue for course-related collaboration. Students are publishing course-related materials within their personal sites on home.uchicago.edu and referencing those materials from Chalk courses and other project sites. These materials are subject to the normal backup procedures of Home, but are typically not considered within the realm of academic content and are seen as personal material even though the content may be referenced or integrated into courses or course-related activities.

4.4.3 Course Material: Concerns

Learning Management

Using the last three complete academic years as a guide, one can estimate Chalk usage in AY2008/2009 if one accepts the following assumptions:

1. the storage architecture of Chalk does not change radically within the next four years
2. the annual increase in new courses slows over time from 18 percent to 5 percent in 2008
3. the average growth in course content slows linearly by 2 percent per year to 37 percent in 2008

Therefore, one can anticipate approximately 1,800 courses and 242 GB of content hosted on Chalk in the 2008/2009 academic year, and a total of 1.2 TB of data for nearly 11,000 courses hosted between AY2001/2002 and AY2008/2009. If trends continue through AY2011/2012 (one decade after the start of formal statistics gathering on Chalk), nearly 3.7TB of content will be stored for the projected 16,000 courses hosted on Chalk over the decade.

The storage implications are obvious, but there are serious policy concerns as well. As of July 2004, Chalk has become the *de facto* content repository for nearly 3,000 courses and a record of student course activity for virtually every student at the University of Chicago. These two facts raise several policy-related questions:

1. When content is published onto Chalk, who owns the rights to the intellectual property? Are the rights owned or licensed by the University for a fixed a period of time?
2. In a related question, if a team collaborates to produce materials for a course or course program such as *American Civilization*, who owns the intellectual property in that case?
3. When students publish course-related content onto Chalk or participate in collaborative activities such as discussion boards and chat rooms, who owns the rights to the intellectual property in those cases? Can that content be archived?
4. Under what circumstances can course content be reused?
5. Given the ability to record and store student course-related activity, what are the privacy issues regarding archival of course materials and course activity?
6. How do Federal acts such as TEACH or DMCA, or Fair Use under U.S. Copyright Law affect the archiving and possible reuse of course content?

Beyond the policy questions are issues surrounding courses stored over time and the retrieval of materials in the future. As projected through 2011, over 16,000 courses will be archived in some manner on Chalk. Aside from the course number, there are few clues as to what is located within an archived course. Courses are actually containers for other pieces of content and may contain a variety of learning objects. Identifying specific learning objects for reuse will be difficult, if not impossible, as little is known about the content within a course when it is archived. If a specific learning object could be extracted from an archived course, could it be reused at all given the pace of technological change? Therefore, what would be the University's responsibility, if any, for ensuring course content could be reused in the future?

Streaming Media

As bandwidth improves and digital media production becomes commonplace, the demand for higher quality materials will increase quickly. A five-minute clip of less-than-VHS quality MPEG-1 video required 20 MB of storage a few years ago; today that same five-minute clip in DVD-quality MPEG-2 would require 150 MB. As technology continues to evolve and bandwidth increases, we can expect larger files of higher quality and varied media types, so storage requirements will continue to grow. Currently, the storage, retrieval and delivery of streaming media is a single coupled service. Technology will change and such a limited architecture may not be feasible in the long-term. Therefore, an issue is what streaming will become and indications are that the shift is toward a broader notion of digital asset management that decouples delivery from storage and retrieval. Once materials are managed as digital assets, the next issue is the problem of describing the stored content. Accurate metadata will be crucial.

If the use of videotape in teaching is any indication, there will be increased interest in using copyrighted materials delivered off of a central media server. This raises concerns regarding digital rights management, appropriate access, and the ability to archive commercial content, and raises the legal issues surrounding DMCA, TEACH Act, and U.S. and international copyright laws.

Other Servers and Pages

There is an unknown exposure regarding course content hosted on other servers and personal Web pages. Because such content is either unknown or part of other unrelated materials, archiving the course material will be difficult. The issue in these cases is whether or not to archive the content in a special manner or to consider the material as merely Web pages subject to the same backup and archival procedures of other pieces of online institutional content.

A different issue that cannot be overlooked is the problem of course-related student activities occurring on local servers. If student progress is tracked and recorded, what are the privacy issues and how can one enforce them when little is known about the server itself?

4.5.1 Research Data: Summary of Issues

Archiving research data presents a host of particularly significant and complex issues ranging from the technical storage and retrieval problems associated with data sets and analysis, to the ever-changing landscape of Federal regulations and grant requirements.

The latter set of issues, regulations, and requirements may present an ongoing significant legal exposure to the University that could impact the institution's ability to receive future funding.

4.5.2 Research Data: Current state

At the present time, there is no central repository for research data. By and large, the archiving of data and related material is left to the principal investigators and research groups, and some divisions and departments are attempting to grapple with the problem at a local level. Most success has been limited to data sets that have mandated usage and access restrictions such as population and census data. Informal inquiries into the state of research data archiving has revealed that much of the long-term storage is in the form of tape and disk backup, rather than a formal archive.

4.5.3 Research Data: Concerns

Storage & Retrieval Issues

Archiving the volumes of data across multiple research projects is and will continue to be an ongoing technical challenge. On the storage side, the volume of computationally-related research data is growing at an alarming rate. Recent studies have found that over an 18-month period where computational performance doubles (Moore's Law), new genomics research data grows eight-fold. Today's researcher is able to acquire much more experimental data than can be processed, so a large amount of the information will need to be stored, accessed, and analyzed at a later time. Researchers who acquire tens of gigabytes of data from thousands of sensors in a single trial will require terabytes of storage in the near term to store repeated trials of the same experiment. That information will be analyzed over months and years and may be shared with researchers around the globe.

Visual and aural research in disciplines such as the humanities and social sciences present a particularly difficult storage and retrieval challenge. Baseline research-grade video (35 Mbps DV) requires approximately 4.5 MB of storage for each second of material, amounting to nearly 15.5 GB of storage for one hour of footage. A project that requires 100 hours of footage will need 1.5 TB of high-performance storage capable of streaming material at rates greater than 35 Mbps. Locating specific details within the material is a challenge, so archiving the metadata associated with the footage will be critical. Metadata may include time references such as SMPTE (Society of Motion Picture & Television Engineers) timecode, as well as scene and analysis metadata. Much of this information may be stored in datasets external to the audio and video materials, so archiving the media objects will not be enough to ensure the data is usable in the future.

As image processing continues to become more accessible to a broader range of researchers, more research projects will involve digital imaging. It is not uncommon to see humanities digital imaging projects involving thousands of unprocessed raw images of between 50 MB and 100 MB per image and more recently, projects have begun to move into the 500 MB per image range. Processed images are typically two to five times larger, therefore an unprocessed 500 MB image may grow to two or more gigabytes in size after manipulation. In addition, each unprocessed image may result in several derivative images as part of research that, in total, may result in a ten-fold increase in storage needs for the project.

Regulatory & Funding Issues

Related to the explosive growth in data acquisition and deferred analysis are the increasing number of Federal regulations for maintaining, storing, and reusing research data. Granting agencies such as the NSF, NIH, and EPA are requiring that research data be available for extended periods -- 25 years or more -- to allow for later reuse of data or validation of current research. For government-sponsored research in other areas, the Freedom of Information Act can apply to the research data and would be subject to the 25-year retrieval requirement as mandated by law. As a result, many grants now require a formal data archiving process to address these regulatory concerns, and it is widely believed that an institution's ability to archive research data will soon be a significant factor in determining project funding.

Access to research data is also a significant regulatory issue. As mentioned earlier, some datasets have strict access and storage requirements. The type or nature of the data may trigger regulatory issues as well. Data defined by HIPAA (Health Insurance Portability and Accountability Act) or FERPA (Family Educational Rights and Privacy Act) as private will require specific data access and management controls. Other data may be deemed as sensitive information and subject to the Patriot Act, thus triggering strict authorization, access, and usage restrictions. Finally, any archived content may fall subject to future Federal laws and regulations. Recent examples include declassified satellite images becoming reclassified for national security reasons. The concern is how to archive information in a manner that allows future regulatory needs to be addressed without disrupting the entire archive or archival process.

Finally, there is a widely-held belief that a backup is adequate for archival purposes. Unfortunately, this is not the case as tape backup over time is shrouded with issues such as long-term format compatibility, hardware stability, software availability, and media longevity. Backups are typically not refreshed, nor are they periodically tested for data integrity. This backup-as-archive misperception can present a major issue of legal exposure for the University.

Appendix 1: Glossary of Terms

Authorization, Policy, and Management Definitions

- **Digital Archiving:** The program (authorized by mandate) for collecting and managing files for preservation

- **Digital Archives:** The files (defined by policy) that are collected and managed for preservation
- **Archival Repository:** The facilities and staff (operated by management) for storage, administration, and accessing of archived files

---

Appendix 2: Selected Metadata Standards

CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. January 2002. (ISO 14721:2003)
This is the standard reference model for building electronic archives. The reference model is available as PDF, http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf, and as a text document, http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/text/CCSDS-650.0-B-1.txt.

Preserving Digital Information: Final Report and Recommendations (1996)
This document, commissioned by the Commission on Preservation and Access and RLG, is the fundamental background document for why and how to build archives whose purpose is the preservation of digital information. Together with OAIS (above), it serves as the basis for subsequent work by digital libraries in this area.

A Metadata Framework to Support the Preservation of Digital Objects (2002)
This document is useful for gaining a deep understanding of the issues involved in preservation archiving, but as it stands it is too complex to implement, hence the need for PREMIS (PREservation Metadata: Implementation Strategies).

Dublin Core

The Dublin Core is a fifteen-element metadata standard recognized and used internationally. It was designed principally by the library and archives community and its primary application is to describe web content.
http://dublincore.org/index.shtml

ISO/IEC 11179
The ISO/IEC 11179 standard describes the definition, specification and content for data element dictionaries and metadata registries.
http://metadata-stds.org/11179/

IEEE 1484.12.1-2002
The IEEE 1484.12.1-2002 Standard for Learning Object Metadata specifies the conceptual data schema that defines the structure of a metadata instance for a learning object of any type, digital or non-digital, that may be used learning and education.