George Mason University

Building On Our Strengths: Digital Archiving, Preservation and Access

Report of the Digital Archiving, Preservation and Access Task Force

March 25, 2004

Wally Grotophorst, University Libraries, *chair*
Dan Cohen, Center for History and New Media
John Creuziger, Technical Systems Division
William Fleming, University Libraries
Polly Khater, University Libraries
Paul Koda, University Libraries
George Oberle, University Libraries
Lene Palmer, University Libraries
Angela Weaver, University Libraries

Building On Our Strengths:
Digital Archiving, Preservation and Access

Contents:

Executive Summary……………………..…………………3

Introduction……………………..………..………………5

Building On Our Strengths: History and Trends…………...….6

Technical Architecture……….……………………….……………..9

Comparative Product Checklist……………………...……..10

Services…………………………….…….………..…………13

Budget / Administration / Staffing / Equipment……..…………17

Implementation /Testing Plan…………….…………………20

About ECHO / CEOSR and the Pilot…………………..21

Appendices

A. Metadata……………………………………...………22

B. Major Features of DSpace System……………………….30

C. Useful links………………………………………………..32

D. Task Force Charge……………………..……...………33

Executive Summary

The long-term retention of digital materials is an immediate need for University Libraries. There are also unmet needs across the university for reliable, permanent, and accessible storage for digital objects (documents, images, multimedia files, databases, etc.).

We propose to use the technical architecture of an "institutional repository" to meet our digital archiving needs and also offer repository services at varying levels to members of the George Mason University community. We hope to identify grant funding for this project, basing our request for funding both on the value of filling this need for George Mason University and serving as a demonstration project to sites interested in reducing the complexity of installing and maintaining a digital repository system.

Functionality

We believe the repository will provide several important ***services***:

- **Accession and Data Storage** – governed by submission agreements negotiated between the Library and object provider.
- **Digital Object Integrity and Migration** – create policies and procedures to ensure the physical and intellectual integrity of objects in the repository. Work with contributor of object to perform transformative migration where required.
- **Discovery and Access** – Support the identification and retrieval of repository objects. Provide OAIS-compliant metadata for objects in the repository to "expose" these objects to users worldwide (subject to access and retrieval limitations negotiated with object contributors)
- **Education and Outreach Services** – Promote importance of digital preservation, explain policies of digital repository, and provide expert consultation on digital preservation issues.

Tiers of Service

We propose to offer this repository resource to the university using a tier-based model that provides varying levels of support based on the object's attributes.

- **Archived -** Materials of significant and widespread value; complex, normalized metadata; commitment to periodic migration.

- **Preserved** – Materials have enduring value, but not enough to merit significant investment currently; basic metadata, supplied by content submitters; commitment to preserve in current format, but not migrate. The bulk of the repository's content will merit this level of service**.**
- **Stored** - Materials not owned or managed by Mason, but which have long term value to Mason scholarship; mirrors of e-journals, other web sites, datasets, CD-ROMs, working papers, and so on. No commitment to migrate or preserve.

Proof-of-Concept Pilot

We propose development of a DSpace installation using equipment already in place within the Library Systems Office. Our test will involve content from three units on campus. We hope to be able offer a fully functional digital repository to the university community during Fall 2005.

Administration of Service

Recognizing that this digital repository service will require a range of talents, we propose that a Digital Repository Group (DRG) be established for the administration of the service. Membership on this team should be drawn from Systems (technology), Cataloging & Acquisitions (accession issues, metadata, object integrity), Special Collections & Archives (metadata, contributor negotiations, policy guidance), Copyright Office (rights management) and Public Services (outreach, retrieval).

We recommend that once established, the DRG fine-tune our recommended implementation plan, develop a procedure to evaluate the pilot and assist library and university administrators in reaching a decision on implementation of a full-scale repository.

Budget

We can begin our pilot phase with minimal hardware funding (we will use existing equipment). To meet the aggressive timetable we propose for our pilot, we recommend assigning one classified staff member to the project full-time. This person will assist Systems Office staff and serve as coordinator with other members of the DRG during the pilot. We expect to incur costs for programming (java) and travel (site visits) during the latter half of the pilot.

To move from testing to a full-scale repository will require an investment approximately $30,000 in hardware (chiefly storage). At that time we recommend at least one librarian be assigned "liaison" duties for the digital repository service.

Our expectation is that over time the digital repository service will grow and additional staff and follow-on hardware will be required.

Final Recommendation

Establish test DSpace installation on existing equipment. Appoint Digital Repository Group to manage pilot testing and evaluation. Task DRG with developing implementation budget and staffing recommendations based on findings of pilot work. Library Administration makes decision on whether to pursue full-scale implementation by June 2005.

Building On Our Strengths:
Digital Archiving, Preservation and Access

Introduction

Over the course of centuries, libraries have developed and optimized a series of "best-practices" for managing information. During most of this period, there has been a relative balance between the scale of content creation (writing and publishing) and the capabilities of libraries to manage the resulting information flow effectively. The "half-life" of paper in this context has allowed archival decisions and subsequent preservation activities to proceed at a deliberate pace.

We are on the brink of important changes...

Today, just ten years after release of the Mosaic browser, we see that the internet, more particularly the web, has forever changed the way we work, communicate, collaborate, find, use and share information. Expanding democracy (anarchy?) in content creation is quickly transforming the landscape of information and posing a disruptive challenge to the science of librarianship. Digital content is flooding an information management structure designed primarily for books, journal articles and bibliographic citations. When the half-life of a web page may be measured in hours, archival decisions cannot be lingered over.

But more than just quickening our pace, we need to confront and develop responses to some of the fundamental changes we already see occurring. The charge to our task force gives voice to several of these:

- publishing is increasingly moving toward *electronic-only* formats and to remain as the primary agency in the scholarly communication process the library must develop mechanisms to insure that this scholarship endures over time. The library has always had (and needs to continue having) a symbiotic relationship with scholarly publishing.

- the library needs to create an environment where digital objects can become part of the permanent collection, ensuring that they are available for use by future generations of scholars
- the university is concerned with protecting its significant and growing investment in digital assets.

Building On Our Strengths: History and Trends

Managing digital content and making it available to users is not a new service in libraries--in fact it has been occurring routinely over the past twenty years in many libraries. [1] Early efforts concentrated on digitizing existing resources (e.g., scanning paper-based content) and success was often measured by the degree of paper emulation the final product presented. Today, with fundamental changes underway in scholarly communication and with parallel developments transforming scholarly publication, paper emulation is no longer an adequate benchmark or a sufficient goal. We need to move toward a new model, fashioning a service that preserves our central role in the university's scholarly communication system but expands it to encompass the evolution in form and process that we see occurring. Phrased differently, we need to develop systems and services that put a clear emphasis on the *library* in digital libraries—fulfilling our role as preservers and distributors of digital content with the same reliable, long-term commitment we bring to more traditional formats.

Of course, as the future unfolds, digital libraries will need to do more than simply store materials or serve as "print on demand" virtual warehouses. Evolutionary change in both information technology and scholarly communication will require that we build infrastructures that can adapt to function in new ways. In a recent interview published in ACM's *Ubiquity*, Cliff Lynch commented on changes he sees coming to the digital library and suggests at least one way in which we will move beyond concerns with storage and retrieval:

[1] University Libraries completed a digital conversion of master's theses in 1988 as part of an experimental project funded by SCHEV. Information on this project may be found in "Keyless Entry: Building Text Databases using OCR Technology," *Library Hi-Tech*, 7:1, January 1989, pp. 7-15.

> *We're going to see digital collections that are presented and managed in a passive way. They will function similar to a repository where stewardship is the major theme. Then you're going to find access systems layered on top of these, which may be more volatile. They may have shorter lives than the underlying collections. You may see the same collections presented through multiple access systems. These access systems will be not just retrieval tools, but analysis environments in some cases. We'll see a great diversity in these access systems -- what I call "digital libraries" as opposed to digital collections.* [2]

Agreeing with Lynch, we see traditional digital collections extending into the future but we also believe that networking technologies will allow us to exploit opportunities to recombine digital objects in new and interesting ways--if we build systems that avoid locking the content in brittle, predefined structures and contexts.

The infrastructure needed to support this sort of activity will include some things that seem quite ordinary to librarians (e.g., selection, preservation, integrity, management) and some that seem new to our domain: interoperability, rights management, unmediated discovery, and so on.

Institutional Repository

It is important to understand that we are proposing a shift in focus from discrete digital collections to a new infrastructure that supports long-term, reliable storage and preservation of a wide variety of digital objects. We will exploit the promise of networked information to support many different ways to organize and present these objects and do so in ways that encourage and facilitate access. In short, we are proposing to assume an organizational commitment to serve as custodian for the many digital objects created by the university community At this moment, other libraries embarking on similar projects are using the term *institutional repository* to describe this sort of work. Consider this oft-cited definition from the Association of Research Libraries:

> *A university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially*

[2] "Check Out the New Library." *Ubiquity* 4, no. 23 (2003).

> *an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.*[3]

Leading this institutional repository effort is the Scholarly Publishing and Academic Resources Coalition (SPARC), an alliance of academic and research libraries [4] and organizations working to, in their words, "correct market dysfunctions in the scholarly publishing system." A recent SPARC whitepaper ("The Case for Institutional Repositories: A Position Paper") offered two primary rationales for institutional repositories:

- Provide a critical component in reforming the system of scholarly communication--a component that expands access to research, reasserts control over scholarship by the academy, increases competition and reduces the monopoly power of journals, and brings economic relief and heightened relevance to the institutions and libraries that support them; and
- Have the potential to serve as tangible indicators of a university's quality and to demonstrate the scientific, societal, and economic relevance of its research activities, thus increasing the institution's visibility, status, and public value.[5]

We see many parallels between the institutional repository and the service we envision—but feel the terminology needs a minor adjustment. We have settled on the term *digital repository* to differentiate one key aspect of our effort: while the technology and infrastructure is nearly identical to what we believe we require, most institutional repositories are primarily focused on providing alternative venues for scholarly publishing. In our particular implementation, we are not placing our primary focus on e-publishing and thus we want to avoid confusion by opting not to use the term "institutional repository." We recognize that e-publishing may well become our primary activity as the system evolves and to build toward that capability, we will include a working papers project in our initial pilot testing.

[3] "Institutional Repostories: Essential Infrastructure for Scholarship in the Digital Age." *ARL Bimonthly Report*, no. 226 (February, 2003): 1-7

[4] George Mason University is a member of the SPARC alliance.

[5] "The Case for Institutional Repositories: A SPARC Position Paper." Prepared by Raym Crow, SPARC Senior Consultant, August 27, 2002. Available via the web at: http://www.arl.org/sparc/IR/ir.html

In many ways, recommending a digital repository and fleshing out the policies and procedures of such a service is a process of transformation. We're not starting at zero--we have well-established policies and procedures to guide our paper-based repository service (99% of the library's collection). We need to develop a similar realm (with clearly stated operational guidelines) for digital materials. We must also recognize that the percentage of our total collection served by this new system will grow dramatically in the future and account for the fact that the infrastructure supporting these efforts may well require that we expand our skill sets.

Technical Architecture

There are many technical approaches to managing digital content/archives but our reading revealed general agreement with the notion that the best solutions separate the archival storage of bits from data management, logical representation and higher-level services. We concur and that general architectural feature is our first requirement. While not comprehensive, here is a skeletal listing of additional requirements we brought to our identification and evaluation process:

- web accessible
- scalable
- Able to handle a variety of bit streams (ASCII text files, Postscript, PDF, Rich Text Format, GIF, JPEG, MPEG, etc.)
- offer a system of persistent identifiers
- able to communicate with metadata harvesting applications and federated search engines
- able to manage digital rights and licensing requirements for users and content providers
- sustainable
- able to migrate to new technologies as they become available

To these functional requirements, we have added three others:

- built with open-source software
- unix-based platform
- demonstrably in operation in at least two locations

Our bias toward open-source software is driven both by a desire to minimize costs where possible and because we believe having the source code for the system will be important during implementation and subsequent development.[6] There are commercial products that meet many (if not all) of the technical architecture requirements we have identified but our research suggests that university-based, open-source projects are where the most interesting and valuable developments are occurring.

We identified three candidate systems (DSpace, Greenstone and Fedora) that appear to meet our requirements. Each is open-source, unix-based, and capable of operating (at least in a pilot phase) on hardware available within the Library Systems Office.

Comparative Product Checklist

The table below contains product information and a high-level list of functions and features. It is not meant to be an exhaustive, in-depth evaluation of the products-–such a review would be impractical given the time constraints on our task force, the rapidly evolving state of digital library software, and the fact that the three products are quite different from one another with respect to market and technological implementation. Each of the products was downloaded to assist in analysis, but they were not installed or made operational.

How different are these products? Our chart summarizes key areas but a few examples might prove instructive: Fedora uses a relational database (in release 1.2) to cache previously rendered XML data streams for performance reasons. The development team is quickly working on replacing the relational database with an XML-based database. The other products rely more heavily on their relational databases for data storage. Both DSpace and Greenstone strive to deliver digital libraries "out of the box." In contrast, Fedora is simply a tool set for the institution wishing to develop a digital library system. The following chart summarizes some key similarities and differences in each package:

[6] One caveat bears mention, by relying on open source software, we will have to develop or acquire local programming expertise as there is no vendor to contact with system-related issues.

	DSpace	Fedora	Greenstone
Provider(s)	MIT & HP	UVA & Cornell	University of Waikato (NZ) & UNESCO
Open Source	Y	Y	Y
Version	1.1.1	1.2	2.41
License	BSD	Mozilla	GNU
Workflow	Y	N	N
Search and retrieval	Y	Services provided through programmatic API's	MG-1.3
Admin module		Implemented via a fat client	Web-based user management, logging, and technical info
Data Organization	Communities & Collections	?	
Database	PostgreSQL	mySQL	
Data Types	Text, images, audio, video	MIME types	
Versioning	Y	N	
OS	Unix or Linux	Windows, Unix, Linux	Windows, Unix, Linux
Submission Process	Distributed (org)		
Dublin Core metadata schema	Qualified (parts)	Y	
METS compliant	N	An extension of	
Export format	XML		
Persistent Identifiers	Handles (not URLs)	Y (not URLs)	URLs
Preservation Models	Bit and functional		
Security	Enforced through Admin module	Access controls via IP addresses	
Storage formats	Bit stream	XML	
Web Server	Apache/Tomcat	Apache/Tomcat	
Developed in	Java	Java	Perl/CGI some C++
Web Services	N	Y	N
Batch		Y	Y

Note: An excellent guide comparing institutional repository software is available from the Open Society Institute. It focuses on several IR packages that we eliminated from consideration (e.g., Eprints) but the report is nevertheless quite useful. DSpace is one of the software packages reviewed.

A locally cached copy is available on the DAPA website at:

http://silo.gmu.edu/da/readings/OSI_Guide_to_Institutional_Repository_Software_v2.pdf

Appendix B of this report contains a high-level look at the DSpace feature set.

Services

Having a system up and running is really just the beginning and in many ways may prove the easiest milestone to meet. The challenges begin to manifest themselves once the first digital object is ready for inclusion. While it is beyond the scope of this report to delve too deeply into the particulars of the repository's operation, a brief discussion of those activities that continue beyond the initial implementation is very important:

Submission/Dissemination Services: The repository administrators will have to negotiate submission agreements with producers to define how, when and under what terms materials will be accepted into the system as well as agreements that determine how and under what terms materials may be searched and extracted.

Digital Object Integrity Services: The repository administrators will have to insure the physical and intellectual integrity of the deposited materials. Using tools like check-sums and digital signatures the physical integrity can be monitored. More serious will be the challenge of maintaining intellectual integrity—tackling issues like: objects that contain links to other objects in the repository or links in a deposited object that point to objects outside the repository.

Data Migration Integrity Services: Not envisioned as an immediate problem but given the long-term commitments we make upon acceptance of objects in the repository, we will have to insure that objects can me migrated or transformed to new, more efficient and cost-effective technologies. We propose two levels of migration service: *Basic* (we provide for deposit, safe-keeping, and return of exact copies up to the point of transformative migration) and *transformative* (we commit to migration that changes the bits of a digital object while retaining essential information). To illustrate the differences, a basic migration would be refreshing data or replicating it on new storage media. A transformative migration would encompass activities like: converting 8-bit ASCII data to 16-bit UNICODE or converting JPG image files to some as yet unknown format. We propose that transformative data migration is a shared responsibility of the producer and the repository administration.

Education and Outreach Services: These services will promote the importance of digital preservation, explain policies and procedures that govern operation of the repository and provide expert consultation and training on digital preservation issues.

Discover and Dissemination Services: The repository must provide this service so users can find and extract needed materials. We recommend sharing our OAIS-compliant metadata with remote harvesting systems to aid discovery and further recommend that requests for an object be via the unique object identifier that is indexed by the repository**.**

Levels of Service

Our research included (and benefited greatly by) looking at work being done at Duke University, specifically their "Digital Archive @ Duke" project:

http://www.lib.duke.edu/its/diglib/digarchive/

Duke has developed a tiered approach to the selection and management of material in their digital archiving system. These tiers recognize that while an archive/repository may well contain a wide variety of objects, they cluster into certain classes as far as selection, processing and maintenance are concerned.

For Mason, where from the beginning we expect to use the framework of an institutional repository to satisfy a number of digital preservation/archiving needs, this tiered approach offers much. Simply put, we will never be able to define (much less implement) a procrustean approach that addresses the wide variety of demands we expect our system to generate within the university community. By developing a system that can provide multiple levels of service (and by extension, satisfy a variety of needs), we believe we can not only improve our efficiency but also maximize the return on the university's investment. For example, if we use our digital repository to support an online library of working papers for a particular college or department, we need not invest time developing complex metadata.

We envision three levels of service: archived, preserved and stored. As collections and/or individual digital objects enter the system, a support level is determined. Subsequent decisions on processing and management flow from the service level determination. Archived materials receive a high level of processing and ongoing management while items merely "stored" in the system receive much less.

Service Tiers

Archived

- Materials of significant, widespread and lasting value
- Complex, normalized metadata
- Commitment to periodic migration
- Stringent criteria to determine what merits this investment

Preserved

- Materials have enduring value, but not enough to merit significant investment currently
- Basic metadata, supplied by content submitters
- Commitment to preserve in current format, but not migrate
- Most data will be in this category

Stored

- Materials not owned or managed by Mason, but which have long term value to Mason scholarship
- Mirrors of e-journals, other web sites, datasets, CD-ROMs, etc.
- No commitment to migrate or preserve, just to store and provide access to local copy as permitted under copyright law.

Budget / Administration / Staffing / Equipment

We have developed a pilot test and guidelines for administration and staffing of this project. It is impossible to give precise figures in some areas because we expect the pilot to help us refine the requirements and fine-tune proposed solutions.

Nevertheless, we can say with confidence that while the pilot will not require significant hardware or software expenditures, we will have to devote resources to staffing and support to meet the ambitious timetable we propose. There is a strong and direct correlation between the resources we dedicate to this project and speed with which it's invention and subsequent development can occur.

Administration

We recommend creation of a Digital Repository Group (DRG) to provide high-level leadership to this project. Chaired by the Associate University Librarian for Systems, we recommend the following representation for the group:

- Systems Office
  - Technology and programming; system administration
- Special Collections and Archives
  - Negotiation of submission agreements; metadata; policy guidance to insure symmetry with traditional archives
- Acquisitions / Cataloging
  - Metadata; cataloging; accession issues; object integrity
- Public Services
  - Outreach; usability issues
- Copyright Office
  - Rights management
- Consultation agreements with ITU staff in Networking, Enterprise Servers and Database Applications (web-based programming and services)

Staffing

To meet the timetable we have proposed (full scale deployment by summer 2005), we recommend that a classified staff member be assigned to the Digital Repository Group full time as soon as possible once the pilot begins. This individual will be the project manager, working under the direction of the AUL for Library Systems. The position also provide liaison with the DLG and coordinate the activities of those providing “as available” assistance during the pilot.

We feel a dedicated staff member is critical to the success of this project—particularly since we are actually proposing to accomplish our tasks based on as-available contributions from professional staff with other pressing duties and responsibilities.

We also anticipate (based on our study of similar institutions that have implemented DSpace) that we will require the services of a java programmer during our pilot test. Recognizing that other departments within the ITU share a similar need, we would like to explore the possibility of developing some sort of resource sharing with the ITU. We could contribute to funding a percentage of the salary for a java programmer in exchange for a commitment of a block of their time. We can better address this requirement once we begin the pilot but our readings suggest we may require assistance for 10-20 hours a week for several months.

Equipment

We have sufficient equipment to begin the pilot test. During the pilot phase, we will want to acquire a small server to test our intended platform (Apple Xserve and Xserve RAID). This unit will cost approximately $3100.00**.** Once the pilot test is completed, and if a decision is made to embark on an enterprise-level repository, we will require an infusion of hardware. At this moment, the platform that offers the greatest performance for the lowest cost is the Apple Xserve and Xserve RAID array. These storage arrays may be chained together to provide additional storage as required.

Apple Xserve Dual G5	$ 4,500
Xserve RAID array (3.5 TB raw capacity)	$ 9,900
(backup: duplicate array or tape)	$ 10-15K
Console (SysAdmin, RAID management)	$ 2,400

Subtotal: $26000 ~ $32,000

Notes on equipment:

- DSpace has been tested on OS X. We would perform our own tests on this platform before committing to the technology for our production system. We will be able to build a Linux-based system at a similar cost. A Solaris system would be somewhat higher.
- We are eager to determine whether it is possible to build this system using Mac OS X (Server), both to take advantage of the close fit with the relatively inexpensive Xserve RAID storage units and to ease system administration tasks for other sites who may wish to follow our example. As distributed and installed on traditional linux platforms (e.g., RedHat, SuSE, etc), the DSpace package requires significant system administration skills. Our hope is that using Mac OS X will enable us to demonstrate a path to these capabilities that is much easier to reach and to maintain.
- If we find that UltraSCSI disks are required, cost for storage will nearly double. We do not anticipate multiple, simultaneous I/O requests for this system and thus are confident recommending ATA technology. We will also consult with representatives from the ITU's Technology Systems Division on this and related issues.

Budget Summary

This summary assumes pilot is successful and decision to implement system on enterprise level is made around June, 2005. Hardware costs reflect prices as of March, 2004.

	Pilot begins	
June 2004	Purchase small server	$3.1K
August 2004	1 classified staff added to project	$35K
September 2004	Travel costs for DLG (site visits)	$2500
September-December	Java Programmer (wages)	$10-25K
June 2005	Xserve upgrade	$4.5K
June 2005	Xraid (storage device)	$9.9K
September 2005	*Production System Begins*	
August 2006	Xraid (2nd storage device)	$9.9K

Implementation Plan – Pilot and Transition to Full Service

We offer the following implementation plan for establishing a digital repository service for University Libraries.

June-August 2004

Procure small Mac OS X server
Install DSpace software
Install PostgreSQL, Tomcat, Apache
Test installation
Obtain Handle Prefix from CNRI (DOI identifier)

September-October 2004

Convene DRG, define assignments
Finalize metadata scheme and documentation
Develop persistent identifier system
Identify and coordinate working papers demonstration project
Import 100 text documents from ECHO
Test OAIS interface
Test export of documents
Test backup procedure

November-December 2004

Add JPG and GIF images to database
Import 1000 documents from ECHO project
Import sample content from SC&A collections
Test OAIS interface
Test backup/recovery procedures
Review and update metadata schemas
Build smaller capacity test server (Xserve)

January – March 2005

Add sample data from CEOSR project (Hurricane Isabel)
Import SQL databases from ECHO project
Import 15,000 MARC records from Voyager
Review and update metadata procedures
Add digital content from SC&A collections
Test limited access mechanisms on selected documents
Review backup procedures

April-June 2005

Fine-tune workflow procedures
Review and finalize management of service within the library
Finalize implementation of working papers pilot
Develop communication plan to publicize service
Establish relationships with likely university users
Add additional ECHO materials
Explore grant opportunities to extend service
Add additional storage to system (e.g., Xserve RAID)

July 2005

Add content from ECHO, CEOSR
Add complete backup of Mason's MARC records (from Voyager)
Reach decision on full-scale rollout of digital repository service

Fall 2005

Full-scale implementation of service begins.

About ECHO and the Pilot

The ECHO project at the Center for History and New Media collects both "born digital" and digitized materials on the recent history of science and technology via the internet. Ideally, the project would like to store its collections within the GMU digital repository with as little transformation of the original objects as possible, The project would also like to be able to search them for individual collections (e.g., a specific set of documents on the history of Usenet) and individual documents (e.g., a specific photograph from the history of hCG research), as well as pull them back out of the repository simply and quickly. At each stage of the implementation plan the ECHO staff plans to work with the Digital Repository Group to assess the ease of these imperatives related to submission, searching, and access. In particular, the ECHO project is concerned about the metatagging, storage, and searching of non-text digital objects, such as images and video, since the project may have a second phase that involves far more multimedia. We hope that there will be as much automation of these aspects as possible, and will work toward this end during the testing phase. Dan Cohen is the DLG's primary contact within ECHO.

About CEOSR and the Pilot

The Center for Earth Observing and Space Research (CEOSR) provides a focus for research done from satellite platforms. An interdisciplinary research center, CEOSR is closely affiliated with the College of Arts and Sciences (CAS) and School of Information Technology and Engineering (IT&E). Our initial work with CEOSR will involve storing digital objects from their Hurricane Isabel (a NASA sponsored Vaaccess-MAGIC project). Hank Wolf is the DLG's contact within CEOSR.

Appendix A

In the course of producing this report, we developed this study of metadata and its relationship to institutional (or digital) repositories. We include it here as a resource for the Digitial Repository Group and their future work with the system.

Metadata

Metadata (information about data) is a fundamental component of any information system—and a basic fact of life for computer users who may have never given the concept much thought. The teenager preparing to make an mp3 copy of an audio CD appreciates not having to type in song titles which automatically fill the screen…thanks to the metadata supplied by FreeDB and similar services. Reliable data that describes digital data and systems that understand and interoperate on standards-based metadata is critical.

For the purposes of a digital repository, there are two types of metadata—descriptive and administrative. Descriptive metadata is information about the item – creator, title, subject, etc. Administrative metadata is information "regarding how the files were created and stored, intellectual property rights, metadata regarding the original source object from which the digital library object derives, and information regarding the provenance of the files comprising the digital library object." [7]

There are a number of multiple metadata standards to chose from when creating an institutional repository. A few we will look at are Dublin Core, EAD, MODS, ONIX, and RDF. Dublin Core is by far the most prevalent and widely used standard. "The Dublin Core metadata element set is a standard for cross-domain information resource description. There are no fundamental restrictions to the types of resources to which Dublin Core metadata can be assigned." [8] Dublin Core originated from a discussion between several OCLC researchers and computer scientists at a 1994 international conference on the world wide web. This informal discussion led to NCSA (National

[7] METS. Metadata Encoding and Transmission Standards. Available at: www.loc.ogv/standards/mets/METSOverview.html (accessed November 13, 2003).
[8] OCLC. Dublin Core Metadata Element Set. Available at: http://dublincore.org/documents/dces/ (accessed November 24, 2003).

Center for Supercomputing Applications) and OCLC holding a joint workshop to discuss metadata semantics in Dublin, Ohio, in 1995. This workshop had more than 50 people discussing "how a core set of semantics for Web-based resources would be extremely useful for categorizing the Web for easier search and retrieval. They dubbed the result "Dublin Core metadata" based on the location of the workshop."[9] Dublin Core has evolved and grown during the past eight years; the DC-Library Application Profile (LAP) working group has since proposed a profile that defines the use of the DCMES (Dublin Core Metadata Element Set) for libraries along with projects and applications.[10]

EAD is Encoded Archival Description. The EAD Document Type Definition (DTD) is a standard for encoding archival finding aids using the Standard Generalized Markup Language (SGML). The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress (LC) in partnership with the Society of American Archivists.[11] EAD is used primarily for archival collections ….

MODS is Metadata Object Description Schema and was initiated and developed by the Library of Congress. It is a bibliographic standard similar to MARC that is expressed in XML; it is currently the closest thing to replacing MARC in the web environment. MODS is intended to carry selected data from existing MARC 21 records into XML, as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format.[12]

ONIX (Online Information Exchange) is a descriptive standard from the publishing community. It is the international standard for representing and communicating book

[9] OCLC. History of the Dublin Core Metadata Initiative. Available at: http://www.dublincore.org/about/history/ (accessed November 24, 2003).
[10] OCLC. DC-Library Application Profile (DC-LAP). Available at: http://dublincore.org/groups/libraries/library-application-profile.shtml (accessed November 25, 2003).
[11] Library of Congress. Official EAD Version 2002 Web Site. Available at: http://www.loc.gov/ead/ (accessed November 24, 2003).
[12] Library of Congress. Metadata Object Description Schema Official Web Site. Available at: http://www.loc.gov/standards/mods/ (accessed November 24, 2003).

industry product information in electronic form. It is developed and maintained by EDItEur jointly with the Book Industry Communication and the Book Industry Study Group, and with user groups in France, Germany, and the Republic of Korea.[13] It is optimized for simple book descriptions, publisher information, rights, purchasing information, etc. It might serve as a foundation from which a richer bibliographic record could be built.

RDF stands for Resource Description Framework and was produced as part of the World Wide Web Consortium's Metadata Activity. It is a standard for encoding entities and the relationships between those entities in a web environment. A resource (anything that can be described by a URI) can have one or more properties and is described by these properties. The relationships between resources are detailed in XML-encoded statements.

The mitigating factor in deciding which metadata schema to use is to determine which scheme works with an OAIS (Open Archival Information System) system and the Open Archives Initiative Protocol for Metadata Harvesting, along with the management software chosen for the repository. The OAI-PMH mandates the use of Dublin Core; that is, "repositories must be able to return records with metadata expressed in the Dublin Core format, without any qualification." [14] A repository can also use and express other types of metadata in addition to Dublin Core (for instance, a locally devised description system for specific collections). One reason George Mason would want to be OAI-PMH compliant is to take full advantage of the OAIster project, which is a project to "create a collection of freely available, difficult-to-access, academically-oriented digital resources … that are easily searchable by anyone". [15] OAIster makes available institutional

[13] EDItEUR. ONIX for books. Available at: http://www.editeur.org/onix.html (accessed November 24, 2003).

[14] Open Archives Initiative. The Open Archives Initiative Protocol for Metadata Harvesting. Version 2.0 of 2002-06-14. Available at: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm (accessed October 28, 2003).

[15] OAIster Home. Available at: http://www.oiaster.org/o/oaister/ (accessed November 24, 2003).

repositories and collections from 239 institutions. As of November 10, 2003, it contained 1,998,524 records for research materials freely available electronically.

The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. Much of their work is centered on the reference model for an Open Archival Information System (OAIS). OAIS is a conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long term. George Mason needs to create and build their institutional repository using the OAIS model to ensure a framework for long-term digital preservation.

It is apparent that DSpace is the software we will use to store and preserve our institutional repository. DSpace also, like the OAI-PMH, dictates the use of Dublin Core. DSpace uses three different "types" of metadata – descriptive, administrative, and structural. Each item in a collection has a qualified descriptive Dublin Core record. The default configuration shipped with the open source for DSpace uses a derivation of the Dublin Core-Libraries Group Application Profile. [16] Communities and collections also have a brief descriptive record stored in the repository. The administrative metadata includes the preservation, provenance and authorization policy information. "Most of this is held with DSpace's relational DMS schema. Provenance metadata (prose) is stored in Dublin Core records. Additionally, some other administrative metadata (for example, bitstream byte sizes and MIME types) is replicated in Dublin Core records so that it is easily accessible outside of DSpace, for example via the OAI protocol." [17] Structural metadata in DSpace is fairly basic; it includes information on how to display an item (in conjunction with all the item parts), along with the relationships between constituent parts of the item. It is noted that the use and storage of structural metadata will be an area of future exploration and improvement in DSpace. [18]

[16] "The DSpace institutional digital repository system: current functionality." Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries, 2003. Pages 87-97.

[17] "The DSpace institutional digital repository system: current functionality." Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries, 2003. Pages 87-97.

[18] "The DSpace institutional digital repository system: current functionality." Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries, 2003. Pages 87-97.

As DSpace evolves, the need is being recognized to incorporate and/or allow other more enriched metadata schemas to be used in conjunction with Dublin Core. Several projects are currently underway to enhance DSpace's ability to interact with and manage items using multiple, unlike types of metadata. SIMILE is one such project; it is supported by the W3C, HP, MIT Libraries, and MIT's Lab for Computer Science. Simile hopes to augment DSpace's support for arbitrary schemas and metadata, primarily though the application of RDF and semantic web techniques.[19] Several institutions have found throughout their applications of Dublin Core, DSpace, and OAI-PMH that Dublin Core does not always enable enough information about items and collections. Dublin Core also tends to be underutilized – elements are not fully and consistently used. [20]

Even with the concern that Dublin Core might not always provide enough avenues for item description, it is the metadata standard that George Mason should use in the institutional repository. It currently is the standard that DSpace and OAI-PMH requires users to employ. It is also an industry norm – which will facilitate our initial foray into the world of institutional repositories and our use of such resources as OAIster.

[19] SIMILE: Semantic Interoperability of Metadata and Information in unLike Environments. Available at: http://web.mit.edu/simile/www/ (accessed November 25, 2003).

[20] Ward, Jewel. "A quantitative analysis of unqualified Dublin core metadata element set usages within data providers registered with the open archives initiative." Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries, 2003. Pages 315-317.

Metadata - Bibliography/Sources Used

- Bainbridge, David, John Thompson and Ian H. Witten. “Assembling and enriching digital library collections.” Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries, 2003. Pages 323-334.

- Chang, Sheau-Hwang. “Institutional repositories : the library’s new role.” OCLC Systems & Services, v.19, no.3 2003. Pages 77-79.

- “DSpace: an open source dynamic digital repository.” D-Lib Magazine, v.9:no.1. Available at: http://www.dlib.org/dlib/january03/smith/01smith.html (accessed November 25, 2003).

- “The DSpace institutional digital repository system: current functionality.” Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries, 2003. Pages 87-97.

- Dushay, Naomi. “Localizing experience of digital content via structural metadata.” Proceedings of the second ACM/IEEE-CS Joint Conference on Digital Libraries, 2002. Pages. 244-252.

- EDItEUR. ONIX for books. Available at: http://www.editeur.org/onix.html (accessed November 24, 2003).

- Gill, Tony. Metadata and the World Wide Web. Available at: http://www.getty.edu/research/conducting_research/standards/intrometadata/2_articles/gill/index.html (accessed November 19, 2003).

- Gilliland-Swetland, Anne J. Setting the sate. Available at: http://www.getty.edu/research/conducting_research/standards/intrometadata/2_articles/index.html (accessed November 19, 2003).

- “Integrating harvesting into digital library content.” Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries, 2003. Pages 183-184.

- Library of Congress. Metadata Encoding and Transmission Standards. Available at: www.loc.ogv/standards/mets/METSOverview.html (accessed November 13, 2003).

- Library of Congress. Metadata Object Description Schema Official Web Site. Available at: http://www.loc.gov/standards/mods/ (accessed November 24, 2003).

- Library of Congress. Official EAD Version 2002 Web Site. Available at: http://www.loc.gov/ead/ (accessed November 24, 2003).

- Medeiros, Norm. "A pioneering spirit : using administrative metadata to manage electronic resources." OCLC Systems & Services, v.19, no.3 2003. Pages 86-88.

- OAIster Home. Available at: http://www.oiaster.org/o/oaister/ (accessed November 24, 2003).

- OCLC. DC-Library Application Profile (DC-LAP). Available at: http://dublincore.org/groups/libraries/library-application-profile.shtml (accessed November 25, 2003).

- OCLC. Dublin Core Metadata Element Set. Available at: http://dublincore.org/documents/dces/ (accessed November 24, 2003).

- OCLC. History of the Dublin Core Metadata Initiative. Available at: http://www.dublincore.org/about/history/ (accessed November 24, 2003).

- Open Archives Initiative. Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting – Guidelines for Repository Implementers. Available at: www.openarchives.org/OAI/2.0/guidelines-repository.htm (accessed October 28, 2003).

- Open Archives Initiative. The Open Archives Initiative Protocol for Metadata Harvesting. Version 2.0 of 2002-06-14. Available at: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm (accessed October 28, 2003).

- Open Society Institute. A Guide to Institutional Repository Software. 1st edition. New York, New York : Open Society Institute, 2003.

- SIMILE: Semantic Interoperability of Metadata and Information in unLike Environments. Available at: http://web.mit.edu/simile/www/ (accessed November 25, 2003).

- Smith, David A., Anne Mahoney, Gregory Crane. "Integrating harvesting into digital library content." Proceedings of the second ACM/IEEE-CS Joint Conference on Digital Libraries, 2002. Pages 183-184.

- Tennant, Roy. Emerging Metadata Standards. PTPL presentation, 2003. Available at: http://escholarship.cdlib.org/rtennant/presentations/2003ptpl/metadata/ (accessed November 13, 2003).

- "University of Washington Early Buddhist Manuscripts Project in DSpace." Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries, 2003. Pages 315-317.

- Ward, Jewel. “A quantitative analysis of unqualified Dublin core metadata element set usages within data providers registered with the open archives initiative.” Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries, 2003. Pages 315-317.

- Woodley, Mary. Crosswalks: the path to universal access? Available at: http://www.getty.edu/research/conducting_research/standards/intrometadata/2_articles/woodley/index.html (accessed November 19, 2003).

Appendix B.

Major features of the DSpace system.

Institutional Repository

DSpace is a digital library system to capture, store, index, preserve, and redistribute the intellectual output of a university's research faculty in digital formats.

- DSpace is organized to accommodate the multidisciplinary and organizational needs of a large institution.
- DSpace provides access to the digital work of the whole institution through one interface.
- DSpace is organized into Communities and Collections, each of which retains its identity within the repository.
- Customization for DSpace communities and collections allows for flexibility in determining policies and workflow.

DSpace accepts any type of digital content, including:

- Text
- Images
- Audio
- Video

Some examples of items that DSpace can accommodate are:

- Documents (e.g., articles, preprints, working papers, technical reports, conference papers)
- Books
- Theses
- Data sets
- Computer programs
- Visual simulations and models

Each institution that implements DSpace can determine its own list of supported formats and content types, based on its needs and resources.

Digital Preservation

One of the primary goals of DSpace is to preserve digital information.

- DSpace provides long-term physical storage and management of digital items in a secure, professionally managed repository including standard operating procedures such as backup, mirroring, refreshing media, and disaster recovery.
- DSpace assigns a persistent identifier to each contributed item to ensure its retrievability far into the future.
- DSpace provides a mechanism for advising content contributors of the preservation support levels they can expect for the files they submit.

Access Control

DSpace allows contributors to limit access to items in DSpace, at both the collection and the individual item level.

Versioning

New versions of previously submitted DSpace items can be added and linked to each other, with or without withdrawal of the older item.

Multiple formats of the same content item can be submitted to DSpace, for example, a TIFF file and a GIF file of the same image.

Search and Retrieval

The DSpace submission process allows for the description of each item using a qualified version of the Dublin Core metadata schema. These descriptions are entered into a relational database, which is used by the search engine to retrieve items.

A more detail specification of DSpace is available at:
http://dspace.org/technology/functionality.pdf

Appendix C.
Useful links

1. University of Oregon:

DSpace implementation at University of Oregon (this installation shows how collections of working papers are handled by DSpace):
http://ir.uoregon.edu

2. D-Lib e-journal

DSpace: An Open Source Dynamic Digital Repository. January, 2003
http://www.dlib.org/dlib/january03/smith/01smith.html

3. DSpace

“Home” site for the DSpace software system
http://dspace.org

To “test drive” MIT’s DSpace:
http://dspace.mit.edu/index.jsp

4. Business Plan for DSpace Installation at MIT

This document describes the business plan developed by MIT for their transforming their research project (DSpace) into a sustainable technology and service. Interesting reading and useful on many levels.
http://libraries.mit.edu/dspace-mit/mit/mellon.pdf

5. University Libraries Digital Archiving, Preservation and Access Task Force (website).

Our task force used this website during our deliberations. It is a useful source of links and documents.

http://silo.gmu.edu/da/

Appendix D.
Task Force Charge

Office of the University Librarian
George Mason University

Memorandum

TO: Digital Archiving, Preservation and Access Task Force

CC: Joy R. Hughes, Vice President for Information Technology & CIO

FROM: John G. Zenelis, University Librarian & Associate Vice President, IT

SUBJECT: Task Force Charge

DATE: August 1, 2003

It is generally recognized that a significant measure of the quality and effectiveness of a research library in the future will be its capability for managing, archiving, preserving, and ensuring access to digital information resources. As we move ever closer to research-level library status, we need to begin planning and developing strategies that ensure that our digital collections -- current and future holdings -- are managed in ways that will assure their availability and accessibility into the future.

The library profession can take justifiable pride in its record as an early developer or adopter of information technology; however, we are now seeing an increase in the pace of change in content as well. The rapid evolution of networked digital information is presenting us with multifaceted challenges, as well as new opportunities. As the nature of digitally-based scholarly information evolves, and the way it is accessed and used continues to change, we need to revisit some of our core competencies -- so that the library may continue to serve as the University's primary agency in the scholarly communication process, an organization that continually meets the research needs of our faculty, students, and staff.

Therefore, we need to begin research and planning for transformational change in the area of digital archiving, preservation and access, to ensure that a digital library of web-based objects and other related components takes root and thrives at George Mason. We need to begin developing both a theoretical and a practical awareness of the issues surrounding digital libraries/archives and begin researching ways to position the library to take advantage of current and future developments.

To assist in formulating plans for our future directions in this vital area, I am asking you to join the Digital Archiving, Preservation and Access Task Force. The scope of the task force's work will encompass: (a) digital collections created by the libraries, (b) Center for

History and New Media (CHNM) selected "born digital" collections which are to become part of the library's permanent collection; (c) digital resources owned; and (d) digital resources "owned" through licenses.

Specifically, the task force is charged with:

- Assess state-of-the art developments in digital resources archiving and preservation. Such review should be designed to insure that our initiatives build on work already done by other institutions. This effort should also identify and recommend adoption of relevant methods and standards in this emergent field (e.g., Open Archives Initiative, Metadata Encoding Transmission Standard, etc.).

- Study options and make recommendations for adopting a defined approach, or range of approaches, addressing "global" rather than specific or targeted projects and/or solutions. This should include the applicability and viability of institution-level initiatives for creating and maintaining "digital repository" platforms such as D-Space, Greenstone, Encompass.

- Recommend an implementation plan for developing a digital library/archive service for George Mason, and make recommendations for the ongoing management of such a program.

- Recommend, plan and provide a cost estimate for a demonstration project or activity – one that will rapidly increase our understanding of the issues, technology tools and management techniques that are already emerging as (promising) standards in this area. This would possibly involve selection of a test-bed system using open-source software whose goal would be to build a working laboratory where we can identify and explore the issues surrounding a digital library/archive. The content for this test system will possibly include materials produced by the CHNM, materials created by our Special Collections and Archives department, and examples of other library owned digital materials.

Wally Grotophorst has agreed to serve as chair of the task force.

I would ask that the task force complete its work and submit its report, including its recommendations, by December 15, 2003.

I thank you in advance for agreeing to serve on the task force and, thereby, for assisting the libraries in this critical endeavor.