UNIVERSITY OF ALABAMA

REPORT ON THE IMAGING SCIENCE & TECHNOLOGY ARCHIVING 2009 CONFERENCE:

PRESERVATION STRATEGIES AND IMAGING TECHNIQUES FOR CULTURAL HERITAGE INSTITUTIONS AND MEMORY ORGANIZATIONS

JODY L DERIDDER

6/14/2009

TABLE OF CONTENTS

Introduction....3

Digital Collection Stewardship....4

- Keynote: From OAIS to DPS to NDHA....4
- Economically Sustainable Digital Preservation....5
- This Library Never Forgets: Preservation, Cooperation, and the Making of Hathitrust Digital Library....5
- Legal Agreements Governing Archiving Partnerships: The NGDA approach....6
- Electronic Records Services for Archival Preservation....7
- Generating Metadata for Digital Preservation: The Chronopolis Scenario....8
- A Selection and Archiving Strategy for Science Records....8
- Preserving Geospatial Data: The national Geospatial Digital Archive's Approach....9
- Assessing the Utility of Current Format Registry Efforts For Geospatial Formats....10
- Meeting the Preservation Demand Responsibly = Lowering the Ingest Bar?....10
- Digital Preservation: Using the Email Account XML Schema....11
- Keynote: Present Status and Next Steps for the Google Book Search Project....11
- One Man's Obsoleteness Is Another Man's Innovation: A Risk Analysis Methodology for Digital Collections....12
- A System for Automated Extraction of Metadata From Scanned Documents Using Layout Recognition and String Pattern Search Models....13
- Barriers to Adopting PREMIS in Cultural Heritage Institutions: An Exploratory Study....13

Imaging and Preservation....14

- Digitising the Dead Sea Scrolls....14
- Preparing for the Image Literate Decade....15
- Metamorfoze Preservation Imaging Guidelines "One Size Fits All"....15
- Summary of the DP3 Project Survey of Digital Print Experience Within Libraries, Archives, and Museums....15
- Management of Spectral Imaging Archives for Scientific Preservation Studies....16
- The Image and the Expert User: A Qualitative Investigation of Decision-Making....16

The Family Search Digital Process.....17

Keynote: Challenges and Opportunities for Digital Stewardship in the Era of Hope and Crisis.....17

Federal Digitization—Moving to Common Guidelines.....18

The Lifecycle of Embedded Image Metadata Within Digital Photographs: Challenges and Best Practices (Or the Secret Life of Photo Metadata).....19

A Status Report on JPEG2000 Implementation for Still Images: The UCONN Survey.....20

From Imaging to Access: Effective Preservation of Legacy Removable Media.....20

Effects on Color Management When Using a Glass Platen to Flatten Book Pages or Documents While Capturing Images With a Digital Still Camera.....21

Implementing Imaging Standards: The Longest Yard.....21

Interactive Presentation: Preparing for the Future as We Build Collections.....22

Conclusions and Recommendations.....23

INTRODUCTION

This sixth annual Archiving Conference , sponsored by the international Imaging Science and Technology organization, was held in Alexandria VA in May 2009. The conference was chaired by William G. LeFurgy, Library of Congress, and divided into two topical sections: Digital Stewardship and Imaging and Preservation. Both topical sections contained panel presentations, preview presentations, and interactive presentations. Each interactive presenter made a short preview presentation before the crowd of almost 200 participants, and then had the opportunity to interact directly with interested parties for over two hours, after the panel presentations and keynotes had finished. There were three keynote speeches, one per day: the first by Steve Knight from the National Library of New Zealand; the second by Dan Clancy of Google; and the third by Clifford Lynch of the Center for Networked Information. Keynotes lasted up to an hour, and most presentations were 20 minutes or less. This venue offered both the opportunity for gathering wide attention and direct feedback, including brainstorming with others about how to leverage new insights and information. The panel presentations will be made available on the website[1], and the papers which were presented are in a bound volume (of which I have a copy; additional copies are available for purchase at a cost of $90 each). The following is a synopsis of the conference presentations I attended, followed by a synopsis of my own interactive presentation and the feedback it elicited. A summary at the end includes my thoughts as to some potential applications and implications for our work here at the University of Alabama.

DIGITAL COLLECTION STEWARDSHIP

[1] http://www.imaging.org/conferences/archiving2009/program.cfm

KEYNOTE: FROM OAIS TO DPS TO NDHA

Presented by Steve Knight, Associate Director National Digital Library, National Library of New Zealand

Digital preservation is not primarily a technical problem. It's a human, social collective organization problem. How can we talk about a "Trusted Digital Repository" as if we are entering into a pact with the future, when we can only guess at the demands that will need to be met? We still have much confusion to overcome. Current tools overlap and are limited in what formats are covered. What gives us confidence that our tools are doing what they should? Various standard developments compete for our attention. How do we know what to trust? We need to agree on the business and functional requirements, as well as the language. We have found that "repository solutions", "digital archiving solutions" and "digital preservation systems" are not likely to be the same thing. We must have economic models for sustaining digital preservation activities. Systemic barriers in many institutions prevent progress, and we must overcome them. Digital preservation must be seamlessly integrated into the organization's business and technology infrastructure. How are we failing to communicate that digital preservation is the biggest challenge facing our institutions now and in the future?

Researchers at the National Library of New Zealand found OAIS[2] unsatisfactory, Digital Preservation Systems better but still unsatisfactory, and are developing our own solution in the National Digital Heritage Archive, in partnership with ExLibris and Sun. However, we as a community need a comprehensive management approach, with strategies to identify and mitigate format obsolescence. Coordinated national/international approaches, a professional services market, and quality assurance standards are necessary next steps. There is a huge gap between theory and practice; the research lags behind. As the data deluge expands exponentially, we also have a capacity capability problem: who is going to do all this work?

ECONOMICALLY SUSTAINABLE DIGITAL PRESERVATION

Presented by Brian Lavoie, OCLC and Fran Berman, San Diego Supercomputer Center

As of 2007, the amount of digital information created, captured, or replicated exceeded the available storage capacity. The problems facing us are technical, social, and economic. The Blue Ribbon Task Force on Sustainable Digital Preservation and Access, a 17-member group of experts from a variety of fields was formed was formed with the 2 year task of framing digital preservation as a sustainable economic activity. The latter will require a deliberate allocation of resources, ongoing or persistent over long periods of time. The interim report of the Blue Ribbon Task Force's effort to articulate the problem and make recommendations was released in December 2008.[3] They have defined economic sustainability, and have determined that it requires: recognition of benefits on the part of decision-makers, incentives for the decision-

---

[2] Consultative Committee for Space Data Systems. 2002. "Reference Model for an Open Archival System (OAIS)." CCSDS 650.0-B-1 Blue Book, January 2002. Available from http://public.ccsds.org/publications/archive/650x0b1.pdf

[3] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. "Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation." Interim Report, December 2008. Available from http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

makers to act, a process of selection of what should be retained, and mechanisms to support ongoing, efficient allocation of resources, and appropriate organization and governance. We need cost/benefit analyses emphasizing the outcomes. We need to identify and leverage institutional self-interest as a business opportunity, mission-driving, and policy compliance. We also need to orchestrate incentives over the entire digital lifecycle.

How do we express the value of digital preservation? We need to talk about the economic threat.

THIS LIBRARY NEVER FORGETS: PRESERVATION, COOPERATION, AND THE MAKING OF HATHITRUST DIGITAL LIBRARY

Presented by Jeremy York, University of Michigan

HathiTrust[4] was launched in October 2008 as a joint undertaking of 25 research libraries to preserve and provide access to millions of volumes of digital holdings. Their focus is content which was converted from print materials, about 16% of which is in the public domain. The Universities of Michigan, California, Virginia, Chicago, Indiana University, and others in the Big Ten[5] have developed a robust and scalable infrastructure based on OAIS and the repository model developed by the University of Michigan to house Google- and locally-scanned content. The effort is funded for five years (2008-2013) by the participating libraries, and includes an executive committee which is considered the "nimble core" of HathiTrust. The current cost model involves estimating the number of volumes to be deposited in a five-year period, averaging the cost per year (which is currently less than 15 cents per volume), plus a one-time fee of 25% of the average annual cost (to help cover one-time resource-intensive needs such as migrating content or changing storage platforms). A new governance and cost model will be developed before the end of the 5-year period, determined by those who have deposited material prior to 2011.

Standards used include METS[6] with PREMIS.[7] Content deposited undergoes ingest validation by GROOVE[8] and periodic MD5[9] verification checks. In addition, a rights database is maintained to associate each object with access and use rules.[10]

---

[4] "HathiTrust: a shared digital repository." http://www.hathitrust.org/about

[5] Michigan State University Administrative Services. "The Big Ten Schools." http://ais.msu.edu/schools.htm

[6] Library of Congress Network Development and MARC Standards Office. 2009. "METS: Metadata Encoding & Transmission Standard." http://www.loc.gov/standards/mets/

[7] PREMIS Editorial Committee. 2009. "PREMIS: Preservation Metadata Maintenance Activity." http://www.loc.gov/standards/premis/

[8] Software developed by Cory Snavely, University of Michigan Library IT Core Services, which uses JHOVE format identification freeware and MD5 checksum scripts.

[9] Walker, John. 2008. "MD5 Command Line Digest Message Utility." http://www.fourmilab.ch/md5/

[10] HathiTrust Permissions agreement, digital assets agreement, rights management, "take down" policy, and access and use policy are all available from http://www.hathitrust.org/rights_management.

"Certification" of volumes in the core collection prevents the need for other institutions to digitize those volumes. Plans are underway to support delivery of born digital content, native XML[11] and encoded text transcriptions, and at some point HathiTrust will expand into audio and video. Services and use of content within the repository will also be expanded. A single search interface does not yet exist, and research is underway to determine if Solr[12] will be able to support full text searching over the entire corpus. One of the services they expect to offer is the ability for certified users at partner libraries to "check out" electronic copies of in-copyright books for use with screen readers and Braille devices.

HathiTrust is a bold effort by leading research libraries to tackle the multiple challenges inherent in developing an enormous collaborative digital library with long-term access support.

LEGAL AGREEMENTS GOVERNING ARCHIVING PARTNERSHIPS: THE NGDA APPROACH

Presented by Julie Sweetkind-Singer and Tracey Erwin, Stanford University; and Mary L. Larsgaard, University of California, Santa Barbara

The Library of Congress National Digital Information Infrastructure & Preservation Program (NDIIPP[13]) funded the formation of the National Geospatial Digital Archive (NGDA[14]) by Stanford University and the University of California, Santa Barbara (UCSB) in 2004. They have developed a set of legal agreements covering authorized users, authorized uses, and management of licensed material. Categories of authorized users include: faculty/staff/students, walk-in patrons, the Library of Congress dark archive, and (in the future) the general public. Authorized uses include those covered by U.S. law, research and educational uses, making multiple copies for preservation, downloading and copying reasonable amounts of content, usage of reasonable amounts within course packs and electronic reserves, and the right to incorporate metadata into a publicly accessible catalog. Custodians are granted rights as well, such as the right to request return of content if it has been deposited elsewhere. Custodians will manage and protect the content, and will remove it from public access within 48 hours in case of copyright infringement. Custodians will also notify users as to terms of use and will credit the copyright holder, as well as identify for the depositor which nodes will hold the content.

The NGDA Content Collection Node Agreement lays out the expectations and obligations of the custodians for long-term retention and use. The accompanying Procedure Manual is a working document (expected to change over time as needed) which specifies the current agreed-upon methods for carrying out these expectations and obligations.

ELECTRONIC RECORDS SERVICES FOR ARCHIVAL PRESERVATION

[11] W3C. Extensible Markup Language. http://www.w3.org/XML/

[12] Apache Solr. http://lucene.apache.org/solr/

[13] Library of Congress. 2009. "Digital Preservation: National Digital Information Infrastructure & Preservation Program." http://www.digitalpreservation.gov/

[14] National Geospatial Digital Archive. http://www.ngda.org/

By Lisa Weber and Hseen Uddin, US National Archives and Records Administration.

The records community and the archival community approach preservation in different ways. In the records community, the aspects of records that determine whether they are to be retained involve their content, context structure, and the significance of the business activity that created them. The context and structure are only captured in the metadata. Management of records involves grouping the records according to hierarchies and disposition plan.

In the archival community, however, the focus is on continuing access; they seek to preserve the content, the context and the structure of the documents. Documents undergo appraisal and selection; the analysis of value is based upon the resources required to preserve the document, not on the inherent value of the document itself.

For records management to become a business process within the archival community, we must determine where the records are being created. Records Management Applications (RMA) are monolithic, cumbersome, costly, complex, difficult to implement and to change. However, the records management community does have components which are helpful to draw upon. For example, the Service Oriented Architecture (SOA) is a business concept. It's a messaging mechanism which relies upon loose coupling.

Within information technology services, the 5015.2 Version I requirements[15] for Electronic Records Management software are not yet proven as the best approach for records management within the archival community. These requirements specify that three elements are used to identify the service to be offered: the bound content (of any standard), the added metadata, and the identification of the item's relationships. The Object Management Group[16] has identified seven services based on the 5015.2 requirements, and has noted a synergy between them and the three elements.

GENERATING METADATA FOR DIGITAL PRESERVATION: THE CHRONOPOLIS SCENARIO

By Ardys Kozbial, Arwen Hutt, and Bradley Westbrook of UCSD; David Minor and Don Sutton, San Diego Supercomputer Center.

Chronopolis[17] is a digital preservation environment built upon a grid-based network, under the auspices of NDIIPP. Each partner operates a grid node containing at least 50 TB of storage capacity for digital collections related to the NDIIPP program. Current partners include the San Diego Supercomputer Center, University of California San Diego Libraries in conjunction with The National Center for Atmospheric Research in Colorado, and the University of Maryland's Institute for Advanced Computer Studies (three nodes currently exist). A subset of the project team, the Metadata Working Group, has traced the path of a digital object through the system to determine what metadata are generated (by system or humans) as it moves through each stage. They have divided this metadata (currently) into eight Event Types, the last three of

[15] DoD 5015.02-STD. "Electronic Records Management Software Applications Design Criteria Standard," April 25, 2007. Available from http://www.dtic.mil/whs/directives/corres/html/501502std.htm

[16] Object Management Group. http://www.omg.org/

[17] Chronopolis Digital Preservation Project. http://chronopolis.sdsc.edu/

which may be repeated. Each Event Type could potentially give rise to another METS record; ingest alone adds about a hundred elements to the metadata. Event Types identified include Service Level Agreement, Acquisition Transfer, Acquisition Validation, Acquisition Registration into SRB[18], Acquisition Registration into ACE[19], Inter-node Inventory Check, Acquisition Replication and File Integrity Check. The Storage Resource Broker (SRB) is an underlying file management system and the Audit Control Environment (ACE) monitors such things as replication. Chronopolis is a dark archive; only those who ingest content can retrieve it.

A SELECTION AND ARCHIVING STRATEGY FOR SCIENCE RECORDS

Presented by John L. Faundeen, U.S. Geological Survey Earth Resources Observation and Science Center

The US Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center archives electronic science collections that currently total over 2 petabytes, and is growing by about a terabyte a day. They have developed an extensive appraisal process for scientific records[20] to help to determine what should be retained, and the costs and challenges for each collection. Thus far, thirty collections have been appraised. Some of the difficulties uncovered include that there is no uniform data model; much of the content is in proprietary formats (some of which are huge); and there are barriers to the data mobility. Some studies cross large time spans, even decades; at what point should content be archived? If content is archived as it is generated, it is quite likely that format migration will have to occur mid-collection.

This experience has clarified for the author an instructive principle, which he calls "the Resurrection Principle." Simply stated, the principle says that: "A preservation system should be resurrectable." Faundeen goes on to say that a preservation system should support its own migration, and should also support some form of handoff of its content.

PRESERVING GEOSPATIAL DATA: THE NATIONAL GEOSPATIAL DIGITAL ARCHIVE'S APPROACH

Presented by Greg Janée, University of California, Santa Barbara

The National Geospatial Digital Archive (NGDA)[21] has begun the development of a format registry[22] to discuss and refine the documentation of geospatial data formats. Some of the characteristics of geospatial data which are relevant to their preservation include: no uniform data model, many formats are proprietary and often huge, there are multiple granularites of data all with their own uses, use is often dependent upon geodatabases (relational databases

[18] Storage Resource Broker. http://www.sdsc.edu/srb/index.php/Main_Page

[19] Audit Control Environment. https://wiki.umiacs.umd.edu/adapt/index.php/Ace:Main

[20] USGS Earth Resources Observation and Science (EROS) Records Appraisal Tool. http://eros.usgs.gov/government/ratool/index.php

[21] The National Geospatial Digital Archive. http://www.ngda.org/

[22] NGDA Format Registry Wiki. http://ngda.library.ucsb.edu/format/index.php/Main_Page

with geographic extensions), and some datasets are continually being reprocessed as new information comes in. These latter must preserve functionality for reprocessing and must track the relationships between related data sets and source datasets as well. In addition, there is both extensive context and implicit context for the studies which must also be captured if the data is to be reused effectively.

Long term preservation can be considered a relay of content over time, a series of hand-offs occurring repeatedly at many levels: between types of media and storage systems, object frameworks and software systems, different institutions and policy regimes, and different communities. As time passes, all links to the original creators and context will be severed, exacerbating the challenges.

The NGDA has identified three architectural design principles beyond OAIS which are necessary to preserve content over time: the "relay" principle, the "fallback" principle, and the "resurrection" principle.

The "relay" principle states that a preservation system should support its own migration. In the event that the system itself is no longer functional, then a preservation system should support some form of hand-off of its content (this is the "fallback" principle). Since the costs of continued migration of content is very high, a method of mitigating that cost is to allow content to become obsolete, but to support sufficient metadata and contextual information to be able to resurrect full access and use at some future time (the "resurrection" principle). Preserving any type of digital information requires preserving the information's context so that it can be interpreted correctly. The NGDA testbed archive system supports all these principles.

ASSESSING THE UTILITY OF CURRENT FORMAT REGISTRY EFFORTS FOR GEOSPATIAL FORMATS

Presented by Nancy Hoebelheinrich, Stanford University

Earlier research has unearthed the assumption that format registries are an important part of digital preservation. We examined the extent to which 23 geospatial data formats and 13 format subtypes in key format registries, and also compared format registry data models.

Though PRONOM[23] and GDFR[24] are setting the standard, these registries have poor and incomplete coverage of geospatial data formats at present. PRONOM had less than 1/3 of the formats, and LOC[25] and GDFR (which is still a prototype system) had even fewer. With some modification, the existing data models would be suitable for registering geospatial formats, but how they will be populated is not yet clear.

MEETING THE PRESERVATION DEMAND RESPONSIBLY = LOWERING THE INGEST BAR?

[23] PRONOM. The Technical Registry. http://www.nationalarchives.gov.uk/PRONOM/

[24] Global Digital Format Registry (GDFR). http://wwww.gdfr.info/

[25] Sustainability of Digital Formats: Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats

Andrea Goethals, Harvard University Library

Current long-term preservation archives operating within the OAIS model assume that producers of content can meet the requirements of ingest. The author proposes that these assumptions need to be revisited.

The Harvard Digital Repository System (DRS) is supporting deposits made by 49 libraries, museums, archives and other units at Harvard, and hosts over 90 TB of content. DRS ingest requirements developed at the outset are becoming barriers to ingest for three primary reasons. Producers are increasingly overwhelmed by the amount of content they need to process for deposit; they have little or no control over the formats or technical properties of the content during creation; and they increasingly need to preserve formats and genres not currently supported by the DRS.

To reduce the barriers to ingest, the Harvard Office for Information Systems is developing a DRS 2, which will support "opaque" objects (of any format), provide minimal preservation services (redundant storage and support of file integrity), and will include new tools and processes to identify files, validate them if possible, extract metadata, and flag problems. For example, the File Information Tool Set (FITS)[26] encompasses JHOVE, Exiftool, the National Library of New Zealand Metadata Extractor, DROID, Ffident, and File Utility, providing a configurable "format tree" and tool ordering preference (since some tools work better for some formats, some for others). FITS will leverage the strengths of these existing open-source tools to identify, validate, and extract technical metadata for a wide variety of file formats.

Batchbuilder[27], the current software used to ingest content, will be expanded to use FITS, and will generate METS objective descriptive files containing PREMIS and more.

Currently, there are no delivery services for "opaque" objects, and they will be accessible only via the administrative interface, as rights issues may not have been resolved prior to ingest. However, this does provide some alternative to complete lack of preservation for these items, and it is expected that some access services will evolve.

DIGITAL PRESERVATION: USING THE EMAIL ACCOUNT XML SCHEMA

Riccardo Ferrante and Lynda Schmitz Fuhrig, Smithsonian Institution Archives.

The Collaborative Electronic Records Project (CERP)[28] was a 3-year pilot conducted by the Smithsonian Institution Archives and the Rockefeller Archive Center to explore the preservation challenges posed by email collections. In collaboration with the Email Collaborative Initiative (EMCAP)[29] they developed an XML schema[30] capable of encompassing an entire email account and its content in a single XML file. The schema structure supports various levels of

[26] File Information Tool Set (FITS). http://code.google.com/p/fits/

[27] Office for Information Services, Harvard University Library. "DRS Batch Builder." http://hul.harvard.edu/ois/systems/drs/bb.html

[28] The Collaborative Electronic Records Project. http://siarchives.si.edu/cerp/

[29] The Electronic Mail Capture and Preservation Project, an NHPRC-funded collaboration between Kentucky, North Carolina, and Pennsylvania. http://www.records.ncdcr.gov/emailpreservation/

granularity which will help to expose social networks and message interrelationships within and across accounts. This schema is in alignment with the email message standard RFC 2822 and has been successfully incorporated into two separately developed email preservation software applications.

CERP decided to archive email as accounts rather than as individual messages, in order to leverage the available resources to best effect, and to retain the original order. XML's nested tagging structure lends itself well to the organization and structure inherent in an email account. The work flow includes a virus scan, making a backup, preservation assessment and analysis and format identification of attachments, development of a finding aid, conversion to MBOX format, parsing the MBOX file and validating the XML output, and creating a METS file to include the metadata and finding aid with the content in a zip file for deposit.

KEYNOTE: PRESENT STATUS AND NEXT STEPS FOR THE GOOGLE BOOK SEARCH PROJECT

Dan Clancy, Google

Of the books included thus far in the Google Books Project, less than 20% are in the public domain. 5% are in print, and 75% or more have an unclear copyright status. Over a 30-day period, users preview at least one page of 78% of the public domain books, and 81% of the partner books. They also view at least 10 pages of 55% of the public domain books in the same time period. Thus inclusion in Google Books enables much wider access to content than could otherwise be achieved.

Google is currently using a remarkably good open source OCR software[31] (they have modified their own to be even better) to convert images to text (in structured XML) to make the content accessible for iPhone and Android phones.[32] Google's "structure extraction" includes graphics and images. They are adding Viewport Syndication and Data API, to enable users to add books to their website, using a widget, to provide content at the locations needed. Google is developing their own scanning technology so they can scale. The tradeoff is between the desired scan quality and the confidence of getting that on each page. The lower the confidence, the lower the cost. It is cheaper to rescan, once a problem is found, than to prevent the problems.

Google is now investing in digitizing historical news, and have formed partnerships with ProQuest and Heritage. For in-copyright material they are approaching the rights holders. For public domain content, they are pulling the content from microfilm. They've developed software to perform block and article segmentation, headlines and advertisement detection.

---

[30] XML Schema for a Single E-Mail Account. September 2008. http://www.archives.ncdcr.gov/mail-account

[31] Tesseract OCR software. http://code.google.com/p/tesseract-ocr/

[32] Examples of Google Books content available in structured text for hand-held browsers: http://books.google.com/m

The settlement agreement[33] that was signed applies to books only, with usage within the United States. It excludes journals, images, and musical compositions. Right now you can preview up to 20%, purchase online access, and obtain institutional subscriptions. 63% of the proceeds go to the rights holder. The settlement agreement includes the establishment of an independent, non-profit Book Rights Registry, which would be a database of rights holders, transferring the 63% proceeds to the rightful owner. If the copyright owner is unknown, this entity holds incoming money up to 5 years while trying to locate the rights holder. Anyone will be able to search this registry to identify rights for in-copyright books.

Google Books Project is dedicated to creating a research corpus containing all the books they scan, which will be licensed to libraries *only* to further research.

ONE MAN'S OBSOLETENESS IS ANOTHER MAN'S INNOVATION: A RISK ANALYSIS METHODOLOGY FOR DIGITAL COLLECTIONS

Kevin De Vorsey and Peter McKinney, National Library of New Zealand

The National Library of New Zealand 's National Digital Heritage Archive (NDHA) is developing a methodology for risk-assessment for long-term access of digital objects, using the Ex Libris Rosetta system for storage . Instead of trying to determine the significant properties of an object, they instead are focusing on identifying and assessing the problematic characteristics of each format. Several problems are not yet solved. For example, at what point is a format obsolete? Is it at the point at which a single character of the document is not renderable in a current software? What if the gist of the content can still be viewed in a current rendering system? The determination of the point of obsoleteness is not yet clear. The NHDA is accepting a definition of "obsoleteness" as the point at which the content is irrenderable by the software at their disposal.

The NDHA is developing libraries that will contain information on formats, their versions, particular characteristics within those versions which are not supported, applications which can render variations of formats, versions and characteristics, and the sustainability factors of both applications and formats. However, business decisions will have to be made about the support of the rendering applications, and these will likely be dependent up on the contract dates that their technical services group has with vendors. As contracts come to an end, migration of supported formats must be considered. This rather practical approach to digital content preservation focuses on supporting a wide variety of formats for the potentially short term during which the current software renders each format. Contract dates provide deadlines for review of content support and decisions for potential preservation action to select methods for continued rendering of impacted formats.

A SYSTEM FOR AUTOMATED EXTRACTION OF METADATA FROM SCANNED DOCUMENTS USING LAYOUT RECOGNITION AND STRING PATTERN SEARCH MODELS

Dharitri Misra, Siyuan Chen, and George R. Thoma, National Library of Medicine

---

[33] Google Book Search Settlement Agreement. http://books.google.com/googlebooks/agreement/

Certain types of textual documents such as journal articles, pamphlets, and official government records contain context-sensitive metadata that can be extracted to assist in search and retrieval of the archived items. The National Library of Medicine (NLM) has developed an automated metadata extraction (AME) system that uses layout classification and recognition models with pattern-matching to extract information from structured and semi-structured textual documents. Thus far it is successful in extracting metadata with 90% accuracy and provides indications when it fails. The system can be customized for similar content, and further enhancements are planned.

BARRIERS TO ADOPTING PREMIS IN CULTURAL HERITAGE INSTITUTIONS: AN EXPLORATORY STUDY

Daniel Gelaw Alemneh, University of North Texas

Most agree that metadata plays a fundamental role in digital preservation. As PREMIS (Preservation Metadata Implementation Strategies) has been extremely influential in providing a "core" set of metadata elements to support digital preservation, this author developed a survey to determine the factors that impact adoption of PREMIS in cultural heritage institutions. Participants included 47 USA institutions (38.2%) and 76 institutions in other countries (61.8%). Institution types included higher education (40%), national libraries (9%), archives (18%), museums (16%), and others. The levels of adoption were divided into 5 stages: "not yet considered," "investigating," "planning," "development," and "adopted." The data analysis revealed that the vast majority of the institutions surveyed had not yet reached the development stage of PREMIS adoption. Of the higher ed institutions surveyed, 20 out of 23 had at least made the decision to adopt PREMIS.

Institutional readiness was shown to be a powerful factor significantly impacting adoption. Several interviewees indicated that while PREMIS can inform their decisions, an institution's specific characteristics must be considered in how or to what extent PREMIS is adopted. Eight factors were most frequently identified which impact adoption, the top four of which follow, with the percentage of respondents who made this selection: "Lack of training/expertise" (48.1%); "Lack of integration or incompatibility with existing system" (37%); "We lack the knowledge necessary to be confident in our ability to implement the PREMIS" (29.6%); "Lack of interest form the decision-makers within our institute" (24.1%). The other four factors include lack of evidence of PREMIS' effectiveness,the perception that the usability requirements are too high, a preference for a "wait-and-see" approach, and a limited capacity to absorb negative consequences that might occur from adoption.

IMAGING AND PRESERVATION

DIGITISING THE DEAD SEA SCROLLS

Presented by Simon Tanner, Kings College, London

The Dead Sea Scrolls were discovered about 60 years ago in caves in the Judean desert. Despite rumors, the scrolls were not in pots; rather, they were strewn about on the floors of 11 caves; many were dessicated and eaten. Approximately 900 scrolls, approximately 2000

years old, have been identified. They now consist of many thousands of fragments which have to be arranged, like puzzles, to reconstruct the scrolls. The Israeli Antiquities Authority asked a number of experts to explore the potential of digitization for conservation, which led to a 2-week pilot in August 2008.

The last full set of images made of these scrolls was approximately 50 years ago, and this included infra-red photography of most of the fragments. At that time the scrolls were laid out on long tables in a brightly lit location (which is damaging) and some fragments were taped together (also severely damaging). Thousands of the fragments had been arranged in plates (between sheets of glass) according to the cave in which they were discovered. Unfortunately, on at least 10 of the 2000 plates, the scrolls then adhered to the glass.

They tested various options and made recommendations for a 3-year digitization project. Text was revealed in Infra-Red imaging that had not been previously recorded, showing the potential impact of IR digital imaging for scholarly discovery. They also experimented with image spectroscopy. The conservation methodology would be to image fragments regularly and use both colorimetry and spectral changes to detect changes before they are visible to the human eye. Changes in parchment reflectance degrade the text legibility and reflect physical changes in the scroll. The X-Rite provided linear Color Checker target proved to be successful in delivering accurate color management with a minimal footprint in the image output. Further experimentation with monitoring water content could become a valuable non-invasive testing mechanism for other manuscript collections.

PREPARING FOR THE IMAGE LITERATE DECADE

Don Williams, Image Science Associates, and Peter D. Burns, Carestream Health, Inc.

Image literacy is emerging: the ability to measure, test, and visually recognize good images from bad ones, based on project requirements. Several initiatives are underway to influence the requirements, using the approach of presenting tools and methods for quantifying and maintaining performance consistency. A successful quality-assurance program will include establishing performance goals, efficient test plans, and performance tracking tools and corrective action.

METAMORFOZE PRESERVATION IMAGING GUIDELINES “ONE SIZE FITS ALL”

Hans van Dormolen, National Library of the Netherlands

Metamorphoze is the Dutch national program for preserving paper originals that are threatened by autonomous decay, cased by acidification, ink corrosion or copper corrosion. The draft version of the guidelines developed to replace the originals with digital derivatives has now been replaced by official guidelines for three different types of originals: those which can be considered works of art, unique library and archival materials, and non-unique library and archival materials. The basic principle is that everything that is visually perceptible in the originals must be also perceptible in the digital version, with the same contrast ratio. Guidelines for originals smaller than DIN A-4 (21 cm x 29,7 cm) and bigger than DIN A-4 (42 cm x 59,4 cm) are still under construction. Current guidelines include calibration of equipment,

monitoring and stabilizing the productions process, color space eciRGBv2, technical tolerances criteria, exposure specifications, contrast levels in the highlights (highlight gamma), and a workflow to create a correction curve to test calibration.

SUMMARY OF THE DP3 PROJECT SURVEY OF DIGITAL PRINT EXPERIENCE WITHIN LIBRARIES, ARCHIVES, AND MUSEUMS

Daniel Burge and Douglas Nishimura, Image Permanence Institute at the Rochester Institute of Technology; and Mirasol Estrada, George Eastman House

The Image Permanence Institute (IPI) conducted an online survey to quantify the experience cultural heritage institutions are having with digitally printed materials. For the purposes of this project, the questions focused on prints created using the most common, non-impact printing technologies: ink jet, dye diffusion thermal transfer and electrophotography. Several types of institutions responded to the survey, 31% of which were libraries, 17% archives, and 25% museums. 71% of the respondents have seen evidence of deterioration in at least some of their digital prints, and 71% do not have specific care policies for these materials. Only 24% of the respondents said they could identify all three kinds of digital prints. Conclusions are that identification training needs to be developed, as well as standards for “permanent” digital prints and care and use guidelines.

MANAGEMENT OF SPECTRAL IMAGING ARCHIVES FOR SCIENTIFIC PRESERVATION STUDIES

Doug Emery, Emery IT; Fenella G. France, Library of Congress; Michael B. Toth, R.B. Toth Associates.

The Preservation Directorate at the Library of Congress has developed a hyperspectral imaging system that measures a series of narrow bands of visible and non-visible spectrum, from ultraviolet through infrared. Collected images may be digitally combined to form processed images which are used for precise analyses of a wide range of materials. This method had the ability to characterize, identify, and possibly quantify materials, discriminating between similar compounds present in the document. Thus the result is a “fingerprinting” of a document, high-resolution full color images, and a characterization of compounds within the document. The imaging system is a MegaVision Monochrome E6, 39 megapixel monochrome back and camera, PhotoShoot image capture software, and light-emitting diode (LED) lights at only around 3 LUX. For the hyperspectral imaging metadata, the LC uses the Archimedes Palimpsest Metadata Standard (APMS), which applies a geospatial metadata model to cultural objects. In addition, the Preservation Research and Testing Division (PRDT) is developing an RDF standard based on the XMP model to manage the technical, administrative, and descriptive metadata and provide links to different types of testing and research data. This methodology will enable expansion on the metadata, semantic description, and interchange of preservation reference data and materials.

THE IMAGE AND THE EXPERT USER: A QUALITATIVE INVESTIGATION OF DECISION-MAKING

Paul Conway, University of Michigan

Research on the development of image digital libraries has so far yielded little knowledge about the actual uses of rich visual content. According to Tefko Saracevic, "Users are from Venus and digital libraries are from Mars." Dr. Conway's paper presented initial findings of a study on the decision making strategies that users employ in a large scale image digital library to choose and evaluate digitized photographs for specific projects. The paper established a foundation for the research in the literature on representation and remediation and described the overall methodology of the research project, which includes two-stage semi-structured interviews with seven expert users who represent the spectrum of factors motivating the sophisticated, project-based use of digitized photographs. Initial findings presented in the paper suggested that users of digitized photographs have relatively little interest in or knowledge of the physical properties of original source photographs and place particular emphasis on image descriptions derived directly from the original objects or from associated filing schemes, rather than making use of metadata assigned by catalogers. Dr. Conway also presented three use cases extracted from his in-depth interviews. The use cases provide some early indications about how users vary in their need for high-resolution images, the value they place on visual and physical context, and their need for specialized tools for viewing and image manipulation . Dr. Conway found that expert users invest significant time and energy in building personal information management tools and augmenting image description with information they obtain through their own research. Can we capture this valuable user-created metadata? Dr Conway also found that users place high value on tools that go beyond search to include navigation within and across image collections.

THE FAMILY SEARCH DIGITAL PROCESS

Richard J. Laxman, FamilySearch

Family Search International is a non-profit organization that captures content from archives worldwide and either hosts them or facilitates archive hosting for genealogical research. They have developed a “Digital Pipeline” to capture 40 million images and metadata per year from original records and from scanning microfilm. In the process, they have developed a new microfilm scanning system and an overhead digital camera system (using CCD array 11 and 16 megapixel cameras). The automated method used for determining resolution is via pixels per line segment, and a semi-automated method of focusing the digital camera involves using an algorithm for lining up check marks with a target. To provide access to collections without a complete index to each record, Family Search associates place specific markers at points in the collection which enables users to search by such options as “county” and “year”; this methodology is called “Waypointing.” Family Search offers an internet indexing application for volunteers to use to download content, transcribe it, and upload the transcriptions for indexing, from their homes.

KEYNOTE: CHALLENGES AND OPPORTUNITIES FOR DIGITAL STEWARDSHIP IN THE ERA OF HOPE AND CRISIS

Clifford Lynch, Coalition for Networked Information

The biggest issues we face right now are public policy issues. For one, we need structures to manages stored material. We are making progress with scholarly work, but the record of the broader social culture is being lost, including government and business records. The situation within the commercial culture (movies, TV, the web) is a patchwork picture, tied up with intellectual property rights. There are conflicts between commercial interests and preservationist interests; the NDIIPP effort is trying to find a middle ground. We need to discuss what our principles are about what *should* be retained, and what rights *should* exist?

Brewster Kahle of the Internet Archive is largely responsible for salvaging a record of our culture via online media. It is sad that we are dependent upon an agency outside the cultural heritage community to preserve our cultural heritage.

While there is some progress in retaining government records at the federal level, funding cutbacks are destroying access and preservation at the state, municipal and county level. Within the corporate community, corporate archives are increasingly winding up in the trash as corporate archivists are losing their jobs. If memory organizations happen to hear about the downsizing, they may be able to collect the corporate records. Alternatively, if the company implodes and is going through the bankruptcy process, we need a policy framework that says there's a public interest in these archives. The focus here is not necessarily on retaining the original records, but in making copies to preserve the information for the cultural record. Relatedly, there is a special public interest in preserving the news, and we need to mobilize our efforts in this area, developing preservation interventions for the public interest.

Our archives are used to obtaining personal papers years after they have been created, all in one batch, when the owner dies. However, the digital record must be captured much more quickly, as it disappears. And since copies can be created and disbursed via many mechanisms, valuable materials are now terribly scattered through privately owned commercial spaces, such as Facebook. We simply cannot travel these realms to collect the archival record.

If an event has enough of a social impact, we need some record of what was going on here. In the past decade, we have been trying to do preservation perfectly. The standards we've been developing are important, in order to lay down foundations. At the same time, the scale of content emerging and the level of demand is too huge in comparison with the resources we have available.

Lynch sees two possible approaches: either we can seek perfection and store very little, or we can be sloppy and preserve more, writing off what is simply intractable.

Given our limited resources, we need low cost, scalable methods to keep materials. We need a “benign neglect” model that saves as much as possible for a future where we hope for more resources with which to deal with the content. This is something we must grapple with seriously.

There will be conflicts between what is rapid, cheap and pragmatic versus perfection; but we *need* to address these issues. And as we move forward, we should preserve open source

source code to document the standards we develop, to enable reconstruction of the content that we save.

FEDERAL DIGITIZATION—MOVING TO COMMON GUIDELINES[34]

J. Michael Stelmach and Carl Fleischhauer, Library of Congress

The field of digital conversion, particularly of historic materials, is new and complex. A survey of current practices and guidelines demonstrated a wider variation among libraries, archives and museums. Some aspects are completely unaddressed. A group of agencies formed initially under the auspices of NDIIPP began an effort in 2007 to develop common digitization guidelines..[35] The focus is to develop goals for categories of content, with objectives stated as use cases. There will be different ones for derivatives. The 2004 NARA guidelines are sufficient for derivatives, but not for masters. We want to change the specifications to address non-output forms of content. The concept of the master is as an "informational" file.

Two working groups have been formed: the Still Image Digitization Working Group and the Audio-Visual Digitization Working Group. All work is being done transparently and all software developed will be made available; some is in beta development now. Already the groups have posited that the taxonomy to be considered primary is that of "signal" versus "noise;" this proposal is posted on site for comment.

The Still Image group is building on the Digital Image Conformance Evaluation (DICE) developed at the Library of Congress, and are publishing a recommendation for TIFF headers based on XMP[36]. The Audio-Visual group is building on the 2006 report on "Capturing Analog Sound for Digitization"[37] and the best practices developed in the Indiana-Harvard collaboration "Sound Directions."[38] Current considerations include Material Exchange Format (MXF)[39] wrapping JP200 and uncompressed video. Everyone seems to be using DPX (Digital Moving-Picture Exchange[40]), but it needs improvement.

---

[34]Stelmach, J.M. and Fleischhaur, C. 2009. "Federal Digitization : Moving to Common Guidelines. PowerPoint available from http://www.digitizationguidelines.gov/stillimages/documents/IS&T_MS_CF.pdf; podcast available from http://www.digitizationguidelines.gov/stillimages/documents/Xerox_Collaboration_JPEG2000_Video.wmv .

[35]U.S. Government. "Federal Agencies Digitization Guidelines. http://www.digitizationguidelines.gov/

[36]Adobe's Extensible Metadata Platform. http://www.adobe.com/products/xmp/

[37]Council on Library and Information Resources and Library of Congress. 2006 . "Capturing Analog Sound for Digitization: Report of a Round Table Discussion on of Best Practices for Transferring Analog Discs and Tapes." Available from http://www.clir.org/pubs/reports/pub137/pub137.pdf

[38]Casey, M. and Gordon, B. "Sound Directions: Best Practices for Audio Digitization." Available from http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/sd_bp_07.pdf

[39]Library of Congress. 2008. Material Exchange Format (MXF). http://www.digitalpreservation.gov/formats/fdd/fdd000013.shtml

[40]Library of Congress. 2005. Digital Moving-Picture Exchange (DPX), Version 2.0. http://www.digitalpreservation.gov/formats/fdd/fdd000178.shtml

THE LIFECYCLE OF EMBEDDED IMAGE METADATA WITHIN DIGITAL PHOTOGRAPHS: CHALLENGES AND BEST PRACTICES (OR THE SECRET LIFE OF PHOTO METADATA)

David Riecks, Stock Artists Alliance, and Phil Michel, Library of Congress

Tremendous quantities of digital images are being created, and it is only a matter of time before this deluge begins to make its way into our archives. It would be ideal if metadata is embedded into the images themselves upon capture, which then can be extracted and used within institutional databases to assist in management and discovery. The World Digital Library[41] (which currently represents 32 different nations sharing significant primary materials) is particularly interested in this effort to leverage existing metadata. Commercial alliances have worked together to develop guidelines to coordinate field properties shared between the Exif[42], IPTC[43] and XMP metadata containers. XMP (Extensible Metadata Platform) is a hybrid between XML and RDF[44]. To see what Exif content your images already have, there is a web tool available which will extract and display the embedded metadata;[45] this tool can be added to your browser toolbar for use on images found in any web page. The proposal to standardize metadata schemas across communities and commercially developed software is available from the Stock Artists Alliance Metadata Manifesto site[46,] and the Photo Metadata Project website[47] provides tutorials and promotional outreach to educate image creators.

A STATUS REPORT ON JPEG2000 IMPLEMENTATION FOR STILL IMAGES: THE UCONN SURVEY

David B. Lowe and Michael J. Bennett, University of Connecticut Libraries

Prior to the release of Djatoka[48] (an open source software for dynamic dissemination of JPEG 2000 images) , the University of Connecticut Libraries surveyed 161 respondents (largely from academic research libraries) on the current use and perceived barriers to adoption of the JPEG 2000 standard. JPEG 2000 supports both lossless and visually lossless (mathematically lossy) compression with a storage savings in comparison with the widely accepted TIFF standard. Though JPEG 2000 has been stable for several years, the tenuousness of its support in the community is enough to cause imaging software developers to question the need for continued

[41]"World Digital Library." http://www.wdl.org/

[42]Japan Electronics and and Information Technologies Industries Association. Exchangeable Image File Format (Exif) standard. More information at http://www.exif.org ,

[43]"International Press Telecommunications Council (IPTC) Photo Metadata. http://www.iptc.org/IPTC4XMP/

[44]W3C. Resource Description Framework (RDF). http://www.w3.org/RDF/

[45]Friedl, J. "Jeffrey's Exif Viewer." http://regex.info/exif.cgi

[46]Stock Artists Alliance. 2006. "Metadata Manifesto." http://www.stockartistsalliance.org/metadata-manifesto-1

[47]Stock Artists Alliance. 2009. "Photo Meta Data." http://www.photometadata.org/

[48]Sourceforge.net. 2009. Djatoka. http://sourceforge.net/projects/djatoka

support of the standard. The survey uncovered several perceived drawbacks of JPEG 2000 , such as codec inconsistencies among software venders, migration concerns, disbelief as to its lossless support, and lack of native browser support. With the exception of the latter, the perceived drawbacks are baseless, and better promotion of the standard is recommended.

FROM IMAGING TO ACCESS: EFFECTIVE PRESERVATION OF LEGACY REMOVABLE MEDIA[49]

Kam Woods and Geoffrey Brown, Indiana University

Within the past quarter-century, many research materials have been published on floppy disks and CD-ROMs, including approximately 5000 items published by the Government Printing Office (GPO). These researchers at Indiana University tested methods of creating bit-identical copies of the original media (as ISO images) for storage and export via a searchable interface[50] on a file server. They uncovered serious errors in both the ISO images and in the software available for transferring the images to the system, and have developed intermediary steps and scripts to enable copy without error, and to detect errors where they exist. Major risk factors include a lack of conformance to the ISO 9660 standard in the original media, and errors introduced by the ISO creation software. Without more than one copy of a given CD, it is impossible to guarantee that a given disc contains the correct bits. Approximately 8% of the images created with Windows based software were truncated. The ISO verification tool "isovfy" provided incomplete or incorrect information in many cases.

The web interface they developed for accessing the content uses SWISH-e for indexing and searching , METS XML records drawn from MARCXML, and XSLT style sheets for display, as well as custom binaries that extract files from ISO and IMG (Floppy images) files, written in C and C++.

EFFECTS ON COLOR MANAGEMENT WHEN USING A GLASS PLATEN TO FLATTEN BOOK PAGES OR DOCUMENTS WHILE CAPTURING IMAGES WITH A DIGITAL STILL CAMERA

Paul Howell and Miranda Howard, Western Michigan University Libraries

An increasing number of libraries and archives are digitizing content using scanners which use glass or other transparent material to hold pages or documents flat for scanning. The metal oxides in standard commercially available sheet glass of any type generally imparts some tint or color to the glass itself, potentially having an adverse impact on the actual colors captured in the image. The researchers provide a detailed analysis of three potential platen materials, using controlled lighting, software, procedures and equipment. The best of the three is the Starfire "Ultra Clear" glass, followed (surprisingly enough) by ordinary plate glass. Acrylic sheet is also suitable, though it is more susceptible to scratching and attraction of dust.

---

[49]Woods, K. and Brown, G. 2009. "From Imaging to Access - Effective Preservation of Legacy Removable Media." Available from http://www.cs.indiana.edu/~kamwoods/woodsbrownarch09.pdf

[50]Indiana University Dept. of Computer Science. 2009. "Sudoc Virtualization Project." http://www.cs.indiana.edu/svp/

IMPLEMENTING IMAGING STANDARDS: THE LONGEST YARD

W. Scott Geffert, Center for Digital Imaging Inc.

The challenge users face in using standards is that these practices are increasingly difficult to apply in the field of competitive commercial systems. Standards by default level the playing field, and it is not to competitor's advantage to support their opponents' products. If the user community allows the industry to control standards, it may be impossible to build a worldwide body of consistent, authoritative cultural images. The author's "capture to print" tests for a single museum became a two-year exploration involving museums world-wide. The most surprising discovery was that images that are measurably accurate to the original artworks reproduce poorly in print without optimization. This disparity is what leads people to manually edit digital captures, a highly subjective editing process with many variables. This suggests that there may be better ways to optimize images, saving time and effort while preserving the integrity of the original image. However many DSLR cameras and raw processors do not easily allow users to apply ICC profiling or measured photography. This research is already impacting software updates and industry leaders, and the author envisions a future where digital cameras are required to incorporate an "ISO Mode" which controls for a true objective capture.

INTERACTIVE PRESENTATION: PREPARING FOR THE FUTURE AS WE BUILD COLLECTIONS

Jody DeRidder, University of Alabama Libraries

Few institutions who undertake the development of a digital library have recognize the depth of commitment involved in supporting long-term access to their digitized content. As workflows are developed, critical choices which can impact long-term support are often overlooked. The first few years are often marked by “boutique collections,” in which each metadata scheme, file naming conventions, organization of files, and even formats may vary. Digital rights and terms of use specifications are often haphazard, provenance metadata may not be kept, metadata may include no reference identifier for the archival file, and commonly no documentation of the standards or choices survives in a predictable and accessible location. Recognition of these problems at an early stage enables reconstruction of missing data and prevention of additional loss.

Currently many digital libraries are facing abandonment due to lack of funding, yet their materials are not organized in a fashion which will enable reconstruction of their digital content at a later date. The speed of obsolescence and the unknowns of digital preservation, coupled with the intense need for low-cost, simple, scalable methods to preserve content, demand an intelligent choice of action. The author recommends a standardized file naming, file organization, and documentation methodology which is within the reach of even small institutions without funding for software or access to a programmer. The patterning of file storage reflects the file name itself, providing a transparent and profoundly simple method of communicating to archivists of the future how the files relate to one another and to the stored metadata and documentation. By leveraging the lowest common denominator, the file system itself, it is possible to sidestep most of the issues related to software dependency. Simplicity, scalability, and low cost make this form of storage an option to almost all digital libraries. Multiple copies of the stored files via LOCKSS ensure continued bit-level support.

Feedback received on this presentation was heartening. Two conference participants asked who our granting agency is, and what other institutions are involved in our effort. It is apparent that there is a pressing need for tractable, low-cost, simple methods to, as Clifford Lynch stated, create a "benign neglect model" for digital preservation. It is clear that the proposed recommendations may serve to offer a method to "mothball" thousands of digital libraries worldwide, in the hopes of later resurrection, rather than lose the many hours of hard work and irreplaceable cultural heritage materials that have already been digitized.

The NGDA principles recommending "fallback," "relay," and "resurrection" principles were evident in the suggestions offered for improvement of the proposed model. Suggestions that were made include storing inode information and information about the parameters of the file system (type of partition support needed, for example) at the top level in a plain text or XML manifest. Some information that should be included are then number and type of characters (Unicode? Latin 1?) allowed in file names, and other expectations for support. Another suggestion was to ensure that each level contains a manifest describing what is at the level below it and how these contents relate. Potentially, the entire file system could be zipped up and included with a manifest for submission into a preservation storage system hosted elsewhere. Another person noted that this file organization proposal is completely format agnostic, which suggests that all types of files could be stored in the hopes of reconstruction or emulation at a later date; there are no format restrictions or standards which must be met; this makes the proposal even more appealing to a variety of institutions, by lowering the bar to enable ease of use.

CONCLUSIONS AND RECOMMENDATIONS

The current funding crisis and economy woes have increased the need for scalable, low-cost, simple methods to prepare existing digital content for long-term storage, with the hopes of resurrecting the precious content at a later date when funding and tools are more available. The time is right to refine and promote our storage model (notably with regard to the recommendations above), providing open-source tools if possible to enhance acceptance and implementation. Further, if the methodology can be expanded to include simple, low-cost open source delivery methods such as an OAI[51] repository built upon the model, a search and browse interface, and ready access to both metadata and derivatives for web agents, this model could easily replace many existing commercial systems, and enable continued support for under-funded digital libraries around the world. By combining both the storage and delivery into the same model, duplication of effort is abolished. By bringing content up to the level of the web, it is possible to support new tools and access possibilities for scholars and researchers. Refinement of the proposed model and expansion upon it is highly recommended. Publicizing our offerings would position us as leaders at a time when the smaller and less well-endowed institutions are most in need.

Other research presented at this conference should also be leveraged to the benefit of the University of Alabama Libraries. The increasing prevalence of the use of geospatial metadata to provide greater usability (as evidenced by both Family Search "Waypointing" and the Library of Congress use of the Archimedes Palimpsest Model) recommends a similar component to be incorporated in our own content, minimally georeferencing materials to the extent possible. If

[51] Open Archives Initiative. http://www.openarchives.org/

embedding metadata in archival content using XMP or a related standard can be implemented without additional costs, we should add this to our workflow. Certainly we should closely follow the developments underway at the federal level to standardize digitization. As the National Library of Medicine refines their system for the automated extraction of metadata, we may find it helpful to make use of some portions of their methodology to reduce metadata creation costs for structured textual materials. To reduce storage costs, we should consider changing our archival image format to JPEG2000 . The Harvard File Information Tool Set promises to be particularly useful as a method of gathering technical metadata from our archival files. When their BatchBuilder tool becomes available, which creates METS files from specified content, altering this software to work over our storage directories would greatly assist in enabling clear and standardized expression of file relationships, while encapsulating all metadata for an object into a single well-organized file.

As Digital Services offers its expertise to other areas on campus and begins to manage born-digital content, it behooves us to learn from the records community how to build a business model for sustainable funding for long-term content support. Mahnaz Ghaznavi of National Historic Publications Records (recently from NARA), recommended a careful study of the Australian DIRKS ("Designing and Implementing Records-keeping Systems"[52]) to inform our decisions and process. Should the UA Special Collections and Archives need to consider long-term access and management of content currently on CD-ROMs or floppies, the software and research at Indiana will prove invaluable. Should campus records begin to face storage of university administration email, the Smithsonian's development of an XML schema to manage this will likely be necessary. In addition, use of the Digital Rights Registry required by the Google settlement will inform decisions on digitization and protect us from legal repercussions.

Last but not least, Paul Conway's research recommends that we move away from librarian-applied metadata in favor of gathering expert user-applied metadata and providing browsability for content over searchability. His findings make it clear that users have much to add, and we need to leverage this capability by capturing user tagging and descriptions, which would be then used for search and display. A shift in emphasis from librarian-created metadata and search-first technology, to more user-friendly access modes and user-interaction will not only enable us to provide far more content at far less cost than our current delivery methodology. This change will also involve the users in the development and promotion of our digital libraries. The greater the involvement of our clientele, the more valuable our holdings become to them, expanding our pool of potential donors.

[52]National Archives of Australia. 2009. "DIRKS." http://www.naa.gov.au/records-management/systems/dirks/index.aspx