

UNIVERSITY OF ALABAMA

REPORT ON IPRES 2008: THE FIFTH
INTERNATIONAL CONFERENCE ON PRESERVATION
OF DIGITAL OBJECTS

JOINED UP AND WORKING: TOOLS AND METHODS
FOR DIGITAL PRESERVATION

Jody L DeRidder
10/9/2008

TABLE OF CONTENTS

Introduction.....	3
Modeling Organizational Goals.....	3
Digital Preservation Formats.....	4
Preservation Planning.....	6
Preservation Metadata.....	7
Grid Storage Architecture	9
Service Architecture for Digital Preservation.....	11
Training and Curriculum Development – Global Overview.....	12
Foundations	14
Conclusions	16

INTRODUCTION

The conference was divided into topical sections; the ones I attended included Modeling Organizational Goals, Digital Preservation Formats, Preservation Planning, Preservation Metadata, Grid Storage Architecture, Service Architecture for Digital Preservation, Training and Curriculum Development, and Foundations. All presentations were 20 minutes or shorter, with limited time for questions; the presentations themselves will be available on the website¹ in the next few days, and the papers which were presented are in a bound volume, a copy of which has been delivered to Tom Wilson. The following is a synopsis of the conference presentations I attended, with related information interspersed which was gathered during question and answer sessions, or in conversation outside the presentations. A summarization at the end includes my thoughts as to some potential applications and implications for our work here at the University of Alabama. (For reference, I have footnoted links to the program² and speaker biographies;³ speakers are denoted in the footnotes with an asterisk.)

MODELING ORGANIZATIONAL GOALS

We need to develop digital preservation policies and strategies for our institution, including roles, structure, and support for the goals of the university. The purpose is long term access; the future benefit to scholars and to the university itself is heavily relevant upon our current policies and implementation. “Any digital preservation policy must be framed in terms of the business drivers and strategies of the institution.”⁴

It’s important to look first at the immediate risks: what is being lost already, because we have no policies for collection, organization, retention, migration, and access? The boundaries cross those of current collections to digital content being created by scholars and administrators, which are tomorrow’s special collections, archives, and library materials. To what extent is the culture and research of today being lost, even as we speak? If we have no strategies, policies, and implementation procedures, we as libraries are reducing our future value to both scholars and to the institution of which we are a part.

The actions we take must match the risk, and the significant properties of the material we seek to save must determine the requirements for implementation. The Planets project has mapped out a complete conceptual model to guide choices, development, and understanding of the problem

¹ <http://www.bl.uk/ipres2008/>

² <http://www.bl.uk/ipres2008/programme.html>

³ <http://www.bl.uk/ipres2008/speakers.html>

⁴ Neil Beagrie*, Najla Rhettberg*, and Peter Williams (Charles Beagrie Ltd.). “Digital Preservation: A Subject of No Importance?” iPRES 2008, London. 29 September 2008.

space. Beta versions of software are also available for download⁵ to assist in planning and assessment.⁶

A suggested business model for digital repositories includes a grid which encompasses the management of service, collection, preservation, business, and information technology on one dimension, and the actions “direct, control, and execute” on the other. Within the grid are collection, access and preservation strategies, rights management, the various aspects of ingest, validation, monitoring, delivery and more. Heat maps used to indicate the areas requiring the most immediate attention assist in organizing one’s thinking about the services needed in the organization.⁷

A third set of business models focuses on our identity as “memory organizations” and the incentives offered by competition for our services. Competitive strategies and potential benefits must be considered next to business goals. Apart from technical implementation requirements, stable organizational, legal and financial models have to be developed which spell out the processes and outcomes in business terms.⁸

DIGITAL PRESERVATION FORMATS

At Cornell, the architecture they decided on in 2004 resembles the OAIS architecture.⁹ They chose MARC XML for descriptive metadata, MySQL for management, and aDORe archival storage architecture¹⁰. The latter was chosen over Fedora because they planned to use Microsoft for access. In aDORe, there is a dual-format storage. Data streams are ARC files, METS files with metadata are kept in the database. They recently moved from JPEG to JPEG2000, as they were impressed by the incorporation of metadata and compression with little loss. They found a steering group necessary to develop an understanding of administrative, operational, and maintenance costs, in order to instigate cost sharing. They found copyright issues to be the main issue, and so they use domain restrictions. They want to support print-on-demand, and are collaborating with vendors. While they focus on bitstream preservation, access is the 800-pound gorilla. They are developing a

⁵ <http://www.planets-project.eu/>

⁶ Angela Dappert* and Adam Farquar (The British Library). “Modeling Organisational Goals to Guide Preservation.” iPRES 2008, London, 29 September 2008.

⁷ Raymond Van Diessen* (IBM Netherlands), Barbara Sierman* (National Library of the Netherlands), and Christopher Lee (University of North Carolina). “Component Business Model for Digital Preservation.” iPRES 2008, London, 29 September 2008.

⁸ Tobias Beinert, Suzanne Lang*, Astrid Schoeger (Bavarian State Library, Munich), Uwe Borghoff, Harald Hagel, Michael Minkus*, and Peter Rödiger (University of Federal Armed Forces). “Development of Organizational and Business Models for Long-term Preservation of Digital Objects.” iPRES 2008, London, 29 September 2008.

⁹ <http://public.ccsds.org/publications/archive/650x0b1.pdf>

¹⁰ <http://african.lanl.gov/aDORe/projects/adoreArchive/>

prototype with Portico for books, as well as working with NYU and Florida (DAITSS¹¹) on the interoperability issue between differing systems.¹²

At Indiana, they are working to archive and enable user access to 2900 CD-ROMs published under the Federal Depository Library Program, printed by the US Government Printing Office. There are an additional 14,000 such items in the Indiana University Library, and more than 120,000 such items in WorldCat, so creating enduring access to this content is not an isolated problem. ISO images are actually good for preservation, but problematic for access. They may contain obsolete formats, executables dependent upon explicit mount points, and various other support problems; in addition, the ISO format itself has changed over time. Over 90% of the CDs tested were missing bits, apparently truncated during image creation. They are translating to and from MARC and METS, using Perl to rewrite links to enable browsing the system, and using Open Office “headless” mode to create xml for content. They have created an emulator using VMWare.¹³

JISC PoWR project¹⁴ is looking at the capture and preservation of web resources of the university itself. Their question to the administration is “What is the web equivalent of a printed prospectus for your institution?” They are using the Firefox Piclens extension to produce an interactive gallery of images, and WetPaint wiki during their workshops to gather input and discuss what to preserve. The agreed-upon targets were the resource, the experience, and the ease of access. They agreed with Chris Rusbridge that the terms “long term accessibility” and “usability over time” are far better than “digital preservation” for obtaining institutional buy-in.¹⁵

The developers of JHOVE2 (Stanford, CDL and Portico) believe that Digital Preservation is managing the gap between what you were given and what you need; the situation is only manageable to the extent that it’s quantifiable. This is where characterization of files comes in (JHOVE is long known as useful for identifying file types and extracting technical metadata). They have changed the terminology to make it more general (now it includes identification, validation, feature extraction, and assessment according to local policy rules) and hope to simplify integration of JHOVE2 into existing processes. No longer is a single object equivalent to a single file; complex objects are recognized, with multiple filetypes. This is important as digital objects have become more complex: a tiff may embed other formats; a JP2 may have many files (JPX fragmentation), and an ESRI shapefile has 3 formats, and 3 files. All the software developed will be available via

¹¹ <http://www.fcla.edu/digitalArchive/index.htm>

¹² Oya Rieger*, Bill Kehoe (Cornell University). “Enduring Access to Digitised Books.” iPRES 2008, London, 29 September 2008.

¹³ Kam Woods*, Geoffrey Brown (Indiana University). “Creating Virtual CD-Rom Collections.” iPRES 2008, London, 29 September 2008.

¹⁴ <http://jiscpowr.jiscinvolve.org/>

¹⁵ Brian Kelly*, Marieke Guy (UKOLN, U of Bath), Kevin Ashley, Ed Pinsent, Richard Davis (U of London Computer Centre), and Jordan Hatcher (Opencontentlawyer.com). “Preservation of Web Resources, the JISC PoWR Project.” iPRES 2008, London, 29 September 2008.

Sourceforge. The project still has about 18 months to go, but there is a public information site¹⁶ and two open mailing lists available for those interested.¹⁷

PRESERVATION PLANNING

Emulation is software dependent. It requires emulator migration, stacked emulation, or virtual machine and modular emulation. This is interesting because generally, format migration and emulation are considered mutually exclusive approaches to digital preservation. However, emulation requires migration of a remarkable amount of metadata and software. Other downsides are that emulation requires many steps; users are unlikely to be computer professionals, and many required components are proprietary and cannot be offered over the internet. The strategy for emulation should be determined by the “viewpath” (the path from object to environment); the more viewpaths that can be retained, the less at risk the document is of obsolescence. These folks have already developed an x86 emulator which runs DOS (called Dioscuri and available from SourceForge¹⁸) and are working on a remote access to an emulation service, entitled GRATE (Global Remote Access to Emulation Services), which supports PRONOM integration (PRONOM is a file registry service to identify types of documents¹⁹) and up- and down- load service across the internet.²⁰ This will be integrated into the Planets project software.

Saving content which has no meaning is useless; how do we determine what the significant properties of a document are? These UK researchers define significant properties as those which must be maintained to ensure the documents’ continued access, use, and meaning over time. They outlined criteria for evaluating significant properties based on such things as the stakeholders, the type of resource, legalities and capabilities. They have developed a data dictionary which will become an XML schema for a PRONOM tool to assist in file analysis.

Computer Aided Design engineering models present a particularly complex type of digital object. Knowledge Information and Management (KIM, UKOLN, Univ. of Bath) has assessed the problem space, potential options, and proposed a framework for long-term management of these objects. Product Lifecycle Management Systems have no preservation planning tools and very little

¹⁶ <http://confluence.ucop.edu/display/JHOVE2Info/Home>

¹⁷ Stephen Abrams* (California Digital Library), Sheila Morrissey (Portico), and Tom Cramer (Stanford University). “What? So What?: The Next-Generation JHOVE2 Architecture for Format-Aware Characterization. iPRES 2008, London, 29 September 2008.

¹⁸ <http://dioscuri.sourceforge.net/>

¹⁹ <http://www.nationalarchives.gov.uk/pronom/>

²⁰ Dirk von Suchodoletz* (University of Freiburg) and Jeffrey van der Hoeven (National Library of the Netherlands). “Emulation: From Digital Artifact to Remotely Rendered Environments.” iPRES 2008, London, 29 September 2008.

available information flow channels. The Planets Project and CRIB²¹ offer options. What are really needed are flexible, modular, consistent registries of format characterization, migration services and more. They have developed two proof-of-concept systems. One is Lightweight Models with Multilayered Annotations (LiMMA) which supports light-weight representation of a CAD model, supplemented with layers of XML-encoded information. They also have a simple preservation planning tool called Registry/Repository of Representation Information for Engineering (RRoRIE), which incorporates a registry of format characteristics and a registry of migration software.²²

Console video games present a complex problem for digital preservation efforts. Using the Planets Preservation planning workflow tool PLATO, various strategies were evaluated using a branch decision model. The success of the evaluation was heavily dependent upon the sample records selected. The emulation alternatives had disadvantages in terms of infrastructure characteristics and metadata. The migration approach ranked very well in most categories with the notable, and expected, failure in interactivity. Emulators supporting more than one system are usually modular and platform independent, but lack compatibility for the games which are dependent upon specific operating system capabilities. While all the tested emulators worked to some extent, most are not usable without modification for digital preservation.²³

PRESERVATION METADATA

The Chronopolis Framework is an NDIIPP funded project to develop a demonstration preservation data grid for large heterogenous sets of data (they are not addressing obsolescence). Within the project, California Digital Library has contributed web crawl content; Scripps Institute of Oceanography has Atmospheric/Oceanic data sets; NCSU has geospatial data sets; and Inter-U Consortium for Political and Social Research has Social Science data sets. Chronopolis provides storage, replication across three nodes, audits, monitoring, replacement and dissemination, and addresses threats such as communication errors, human error, malicious attack, hardware/media failure and natural disaster. For each ingested dataset, they identify the processes to be supported, the metadata needed, and coordinate distribution on the network and outputs. Each internal process (ingest, replication, auditing, and dissemination) creates additional metadata, which is stored in 4 locations with some overlap. This is considered a strength: they identify the file, document its history, determine if the requisite metadata exists, and define the relationship between the metadata in its separate locations. PREMIS is used for the bulk of the encoding. The

²¹ <http://whitepapers.techrepublic.com.com/abstract.aspx?docid=287512>

²² Alexander Ball*, Manjula Patel, Lian Ding (UKOLN, University of Bath). "Towards a Curation and Preservation Architecture for CAD Engineering Models." iPRES 2008, London, 29 September 2008.

²³ Mark Guttenbrunner*, Christoph Becker, Andreas Rauber, Carment Kehrberg (Vienna University of Technology). "Evaluating Strategies for Preservation of Console Video Games." iPRES 2008, London, 29 September 2008.

output is a Dissemination Information Package which is used to provide reports to data providers, transfer to another repository, and the file history audit from within the repository itself.²⁴

The British Library has developed a system for ingest, storage, and preservation of digital content using eJournals as the first data stream. They developed a common Archival Ingest Package (AIP) structure for them, after studying the business processes and data structures and using the National Library of Medicine DTD to normalize content files. They created an AIP for each level (article, issue and journal) which enables each to be updated separately. The structural information was separated from the files, and thus wound up with 5 METS records for each object: one for submission, one for the manifestation of all the files needed for dissemination and the relationships of all the METS files, and one for each of the three levels. To keep track of all this, they kept copies of all the metadata in a separate database called the Metadata Management Component, to provide fast access for manipulation. Relationships between items were tracked using MODS relatedItem fields as well as PREMIS relationships. They used PREMIS amdSec, dmdSec and extension schemas, and had 3 identifiers for each object: one for the intellectual identity, one for the generation dependent-related item, and one for the archival record.²⁵

The German “kopal” software is an example of an OAIS system, using the Universal Object Format²⁶ (METS and preservation metadata), developed for web archiving. They have two approaches: selective, or domain specific, and use the Heretrix crawler (a project of the International Internet Preservation Consortium). They are using a newer version of the container format ARC, called WARC, which is a draft ISO standard. The container includes several records, including binary files, information about the web crawl, changes since the last collection, conversion information and metadata. One question that arose was why they should need to use METS when they already have a container (WARC)? In the Minerva project METS was used at the aggregate level and page file level, but in their Web Curator project they instead used METS about the crawl but not about the content.

Migration of web content is high in complexity, particularly with regard to the links. Metadata is needed about the file formats and dependencies. One potential solution is emulation of web browsers of a certain time period, but this requires metadata about the software and operating systems of each successive time period, and dependencies of plugins. This metadata should be included in archival packages in the present system: METS with technical metadata information on files for migration, or METS with crawl information and the content in a WARC for emulation.²⁷

²⁴ Arwen Hutt*, Brad Westbrook, Ardys Kozbial (University of California), Robert McDonald (Indiana University Libraries), Don Sutton (San Diego Super Computer Center). “Developing Preservation Metadata for Use in Grid Based Preservation Systems.” iPRES 2008, London, 29 September 2008.

²⁵ Angela Dappert, Markus Enders* (The British Library). “Using METS, PREMIS and MODS for Archiving eJournals.” iPRES 2008, London, 29 September 2008.

²⁶ http://kopal.langzeitarchivierung.de/downloads/kopal_Universal_Object_Format.pdf

²⁷ Tobias Steinke* (German National Library). “Harvester Results in Digital Preservation System.” iPRES 2008, London, 29 September 2008.

The Library of Congress has been exploring the usefulness of FRBR²⁸ in designing systems for Cultural Heritage. One barrier they ran up against is that of description versus design; a data model that is useful for programmers is not necessarily helpful for those trying to reason theoretically about bibliographic relationships. They developed a "Paper Tool:" a collection of symbolic elements whose construction and manipulation follow rules and constraints of one or more guiding theories. Representations are included for work, expression, manifestation and item, and an item can be a container comprised of more than one item. The author recommends that instead of using FRBR to guide design, we should consider it a resource, about which we can make business assertions. We need to consider how we distinguish between resources.²⁹

GRID STORAGE ARCHITECTURE

The JISC-funded Preserv 2 project³⁰ recognizes that the huge rise in volumes of born-digital content requires automatic processes for selection and evaluation for preservation; manual processes will not scale. They are working to create a model for this as they develop methods for preservation of institutional repositories (IRs). They are currently using the Sun Microsystems STK5800 Honeycombs (which are object oriented), Amazon Simple Storage Service (S3) and SWORD (Simple Web-Service Offering Repository Deposit).³¹ They have a 3-stage approach: format identification and characterization, preservation planning and technology watch, and preservation action and migration. For the format ID services, they are using PRONOM-DROID from the UK. DROID (Digital Record Object Identification³²) is downloadable, and it works with the PRONOM³³ online format registry to both identify the formats and to schedule events and notify by email if an event has taken place (using Calendar, Outlook, and Sunbird). They are adding the Plato preservation planning tool from the Planets project.³⁴ The weaknesses of the automatic system are that we must have registries for each format, regular evaluation of risks to each format and must identify tasks and timing for each migration.³⁵

²⁸ Functional Requirements for Bibliographic Records. <http://www.ifla.org/VII/s13/frbr/frbr.pdf>

²⁹ Ronald Murray (Library of Congress). "The FRBR Theoretical Library: The Role of Conceptual Data Modeling in Cultural Heritage Information System Design." iPRES 2008, London, 29 September 2008.

³⁰ <http://preserv.eprints.org/>

³¹ <http://www.ukoln.ac.uk/repositories/digirep/index/SWORD>

³² <http://droid.sourceforge.net/wiki/index.php/Introduction>

³³ <http://www.nationalarchives.gov.uk/pronom>

³⁴ <http://www.ifs.tuwien.ac.at/dp/plato/>

³⁵ Steve Hitchcock*, David Tarrant, Leslie Carr (University of Southampton), Adrian Brown (The National Archives), Ben O'Steen, Neil Jeffries (Oxford University). "Towards Smart Storage for Repository Preservation Infrastructure." iPRES 2008, London, 30 September 2008.

Keith Rajecki from Sun started his presentation by noting that the Honeycomb system used in the Preserv 2 project above, is tightly coupled to the x2100 server, and there is an EOL (end to life) for both. He briefly described the current systems for sale -- Lustre for clustered file systems separating data management from storage management, and the dollar per gigabyte SATA drives in the storage server cluster Sunfire x4500, offering 4 tiers local and remote in an "infinite archive" where content is transferred to tape archives based on rules we set. However, they are still working to develop platform-agnostic code. The software systems they are offering are still tightly coupled to the hardware; when the hardware is obsolete, so is the software!³⁶

More and more institutions are establishing data grids using the San Diego Supercomputer Center's Storage Resource Broker (SRB) for managing their collections. We need automated methods to monitor file formats and notify managers of content at risk and recommend solutions for long-term access. PresSRB is a test bed implementation for obsolescence detection, notification and migration over a heterogenous distributed collection of objects stored in an SRB. They reused the approach of the PANIC³⁷/AONS³⁸ project. PANIC is a semi-automated preservation system that relies on semantic metadata to provide obsolescence detection, notification and migration. The AONS project aimed to adapt the detection and notification components to generate a web service that could be applied to multiple collection types. They are storing preservation metadata in MCAT.³⁹ Remote sensing satellite images are often proprietary, so they were translated to geoTiffs for preservation. The Geospatial Data Abstraction Library (GDAL) was used for translations of raster formats.⁴⁰

In the SRB, users access objects across a distributed environment, using parallel I/O⁴¹, multithreading and bulk operations; it's very fast. PresSRB prototype uses PHP and command line scripts in Linux with Apache. Linux "file" command is used to identify format (they modified this for BigTiff and ERDAS Imagine files, and to provide native SRB support); and obtains format information from registries to identify obsolete formats. GDAL gdalinfo extracts resolution, layers, geographical metadata, and gdal_translate migrates the content, giving a preview of the results first. Two format registries are used: LCSDF and PRONOM.⁴²

³⁶ Keith Rajecki* (Sun Microsystems Inc.) "Repository and Preservation Storage Architecture." iPRES 2008, London, 30 September 2008.

³⁷ <http://metadata.net/panic/>

³⁸ <http://sourceforge.net/projects/aons>

³⁹ <http://www.sdsc.edu/srb/index.php/MCAT>

⁴⁰ <http://www.gdal.org/>

⁴¹ Input/Output.

⁴² Douglas Kosovic*, and Jane Hunter (University of Queensland). "Implementing Preservation Services over the Storage Resource Broker." iPRES 2008, London, 30 September 2008.

In the discussion that followed this presentation, San Diego Supercomputer reps estimated that it costs a million dollars currently to keep a petabyte online. A BBC rep said that they have millions of audio and video shelved – approximately 10 petabytes, that they simply cannot afford to put online.

The SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) project seeks to create a framework for long-term preservation of digital library content on the data grid. They are embedding legacy content into the grid, using a Daffodil⁴³ user interface and iRODS⁴⁴ search and browse service, and iRODS wrappers (or similar for abstracting legacy delivery systems such as DSPACE and KOPAL.⁴⁵

SERVICE ARCHITECTURE FOR DIGITAL PRESERVATION

DAITSS (Dark Archives In the Sunshine State)⁴⁶ has been running since 2005; it currently handles 11 TB and is absorbing another ¼ TB per week. They have 2 50-TB disarrays (different locations) backed up to tape. It is a monolithic Java program, which is difficult to install and configure. It also has internal coupling required between services, and its SOAP (RPC) interface has turned out to be quite brittle. They are changing over to a REST interface (PUT, GET, DELETE, HEAD –the latter gets simple metadata), and they are seeking to make the components independent. Application states will be divvied up into resources, represented by URLs. There are 5 services in the Ingester - Description, Action Plan, Transformation Service, AIP Service, and the Storage Service. PRONOM is used to identify the format, and the appropriate technical metadata is extracted, using Jhove1 and 2, Ghostscript, ffmpeg library, mencoder and the libquicktime suite. In the Action Plan, links are URLs for transformation services in an XML file, offering normalization, ingestion, hypertext as an engine of application state. In the Transform service, a Request (sendfile) returns the link to the new object. An interesting note here is that SSH will only send up to 2 GB. If the transformation does not work well, there is no backup, as the original is overwritten(!) DAITSS 2 Service architecture will be a distributed web services model.⁴⁷

A conceptual framework for the Service-Oriented Approach towards digital preservation includes 4 basic concepts: a front end, the service itself, a repository service bus, and the need to be loosely-coupled and platform-independent. This framework was developed using Service Component Architecture⁴⁸ (SCA) and the Business Process Execution Language⁴⁹ (BPEL).⁵⁰

⁴³ <http://www.daffodil.de/>

⁴⁴ <https://www.irods.org/>

⁴⁵ Claus-Peter Klas*, Holger Brocks, Lars Müller, Matthias Hemmje (Fern Universität, Hagen). “Embedding Legacy Environments into a Grid-Based Infrastructure.” iPRES 2008, London, 30 September 2008.

⁴⁶ <http://daitss.fcla.edu/>

⁴⁷ Randall Fischer*, Carol Chou, Franco Lassarino (Florida Center for Library Automation). “Updating DAITSS: Transitioning to a Web-Service Architecture.” iPRES 2008, London, 30 September 2008.

⁴⁸ http://www.davidchappell.com/articles/Introducing_SCA.pdf

In Portugal, they have already built a Service-Oriented Digital Repository:” RODA (Repository of Authentic Digital Objects). They used a simplified version of EAD for the descriptive metadata, and CRiB (Conversion and Recommendation of Digital Object Formats) for their distributed migration service (based on a Ph.D. thesis by the speaker). They normalized text, image and relational database content into PDFs, TIFFS, and XML format (for databases). The metadata is stored in a database, the content goes into Fedora. They compared DSpace and Fedora, and though DSpace outperformed Fedora on most of their required needs, it lacks flexibility, expansibility, and support for their chosen descriptive metadata format. They built on the the Fedora preservation metadata and include integrity checks. They have noted that with over 10,000 items, an XML database begins to fail.⁵¹

In the Cultural Heritage domain, it is insufficient to identify a resource with a persistent identifier; we also need to guarantee authenticity, credibility, and continuous access to it. Key to this is the long term sustainability of the Registration Authority who maintains the association register and access. The Nordic Metadata Project has adopted the National Bibliographic Number (NBN, used by the National Libraries in Italy for the development of a trusted digital repository), registered a namespace and developed a prototype for a national register of digital cultural resources. Now they are looking for international cooperation. They want to reinforce the peer-to-peer resolution service and develop a protocol for inter-domains resolution services.⁵²

TRAINING AND CURRICULUM DEVELOPMENT– GLOBAL OVERVIEW

The Digital Preservation Management Workshop, an NEH-supported effort, has been operating for five years now. It was developed at Cornell with Ann Kenney, and is now moving to Michigan. It’s a 5-day workshop based on 3 stool legs of organization, technical infrastructure, and resources, with 5 stages of activities taught: acknowledge, act, consolidate, institutionalize, and externalize. They focus on action plans and gap analyses, and short-term solutions for long-term problems. They also have a 2-day version off-site workshop. The target audience is managers; they are not training them on technical implementation. The breakdown of their audience origins is 52% academic libraries, 25% government, 13% corporations, 9% museums and cultural heritage institutions, and 2% public libraries. They have an online tutorial⁵³ and prerequisite assignments

⁴⁹ http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsbpel

⁵⁰ Christian Saul*, Fanny Klett (Fraunhofer Institute of Digital Media Technology). “Conceptual Framework for the Use of the Service Oriented Architecture Approach.” iPRES 2008, London, 30 September 2008.

⁵¹ Jose Carlos Ramalho*, Miguel Ferrieira (University of Minho), Luis Faria, Rui Castro, Francisco Barbedo, Luis Corujo (Directorate General of Archives in Portugal). “RODA and Crib: A Service Oriented Digital Repository.” iPRES 2008, London, 30 September 2008.

⁵² Emanuele Bellini, Chiara Cirinná, Maurizio Lunghi* (Fountazione Rinascimento Digitale), Ernesto Damiani, Christiano Fugazza (University of Milan). “Persistent Identifier Distributed System for Cultural Heritage of Digital Objects.” iPRES 2008, London, 30 September 2008.

⁵³ http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html

such as performing a readiness survey and reading the OAIS (Open Archival Information System⁵⁴) document.

The top threats are insufficient policies or plans, and insufficient resources for preservation. Attendees explore legal issues, digital content assessment, and (de)constructioning archival objects. More content is being added to the tutorial by May. They have Train-the-Trainer options as well. They have worked with the Digital Preservation Coalition and Digital Preservation Training Programme, and are now working with Data Curation Education Program and the Digital Curation Centre, Virginia Tech, and University of NC to develop a template for documenting curriculum modules. They are developing a “Workshop in a Box” which will probably be unveiled at the DigCCurr 2009 conference in April.⁵⁵

The Digital Preservation Training Programme⁵⁶ was kickstarted by JISC in 2004 with Digital Preservation Coalition and the Cornell training program. Required reading includes the OAIS reference model and TDR (Trusted Digital Repositories⁵⁷). Participants are required to identify who in their organization performs each OAIS function. If too many are being performed by a single person (this is common), there’s a problem. They have developed formulas to determine migration costs versus emulation costs, which look to be very useful. Training is currently 2.5 days on site or in London, and in addition, they are developing an online tutorial. The market is fragmenting, and new positions are showing up, such as: repository manager, scientific data curator, and digital programs director.⁵⁸

Rachel Frick of IMLS says they are the largest federal cultural funding agency, and that through the Library Science and Technology act they provide state funding to support technology and continuing education. They also have competitive grants intended to elevate practice around preservation. They have a project called the Heritage Health Index⁵⁹ which spells out the condition and preservation needs of US collections. Their National Leadership Grants are designed to promote preservation while enhancing learning. They support the UNC Chapel Hill digital curation curriculum development and the UIUC MLS concentration in data curation. At NYU they are involved with the Moving Image Archiving and Preservation studies in preserving video. Other

⁵⁴ <http://public.ccsds.org/publications/archive/650x0b1.pdf>

⁵⁵ Nancy McGovern*, Aprille McKay* (Inter-University Consortium for Political and Social Research.” iPRES 2008, London, 30 September 2008.

⁵⁶ <http://www.ulcc.ac.uk/dptp/>

⁵⁷ <http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf>

⁵⁸ Kevin Ashley* (University of London Computer Centre). “Digital Preservation Training Programme.” iPRES 2008, London, 30 September 2008.

⁵⁹ <http://www.heritagepreservation.org/HHI/>

involvements include the Northeast Document Conservation Center, Safe Sound Archives, and the ICPSR (Interuniversity Consortium for Political and Social Research).⁶⁰

The Digital Curation Centre⁶¹ continues to develop useful free resources and publications to assist and inform efforts in digital preservation. SWOT analyses are needed on all levels (Strengths, Weaknesses, Opportunities, Threats), and perhaps we need to develop competence centers. We need information management modules to address how content will be accessible for the long term. One question which arises is, do we need to be information managers first? Whether our backgrounds are in archival science, information science, or elsewhere, we must understand information, how it is use, and the organizations in which it is used. The NC workshop (DigCCurr 2009⁶²) will focus on how to solve problems. We need 5 year plans, 10 year plans, preservation policies and collaborative initiatives.⁶³

FOUNDATIONS

David Rosenthal began his presentation with the disturbing statement that Sun specifically disclaims all liability for data loss. Rosenthal, now at Stanford, used to work for Sun, Sirius Cybernetics, and Pergamum (UC Santa Cruz). He states that all studies of file systems and RAIDs find serious bugs. Their claims for sustainability are based on a simplistic threat model that completely leaves out operator error, economic or organizational failure, and internal or external attack. Hardware reliability is a hot topic now, as “silent data corruption” has been found to be rampant in state-of-the art equipment. In practice, we are all losing data. What constitutes acceptable loss? Even CD’s replay with errors and minor data loss on a continual basis. We are better off purchasing cheap equipment and making copies across a variety of hardware and geographical area, than spending high dollar amounts on single solutions.⁶⁴

In the discussion that followed, David’s coworker at Stanford stated that LOCKSS does not scale. Once you’re into the multiple petabytes range, the bandwidth will not support it feasibly.

The University of Arizona has developed a modeling process to evaluate the reliability of storage media and backup systems for both physical and logical preservation. Logical preservation is more complex than physical because it requires technology and processes to ensure that bitstreams are

⁶⁰ Rachel Frick* (Institute of Museum and Library Services). “Funding Digital Preservation Research Practice and Education in the US.” iPRES 2008, London, 30 September 2008.

⁶¹ <http://www.dcc.ac.uk/>

⁶² <http://www.ils.unc.edu/digccurr2009/>

⁶³ Jody Davidson* (Digital Curation Centre). “The Key Challenges in Training and Educating a Professional Digital Preservation Workshop.” iPRES 2008, London, 30 September 2008.

⁶⁴ David Rosenthal* (Stanford University). “Bit Preservation: A Solved Problem?” iPRES 2008, London, 30 September 2008.

renderable and accessible. They uncovered that there is a higher deterioration for optic and magnetic media with exposure to high temperatures and humidity.⁶⁵

The “nestor Catalog of Criteria for Trusted Digital Repositories” of 2006⁶⁶ is highly recommended. Standards are growing far faster than digital archives. For example, PREMIS tells us what preservation metadata we need; METS tells us how to package it; PAIMAS (Producer-Archive Interface Methodology Abstract Standard⁶⁷) lists nearly 90 steps for ingest alone, and DRAMBORA (Digital Repository Audit Method Based on Risk Assessment)⁶⁸ enumerates possible preservation risks on over 200 pages. Traditional archivists are not involved in the standards being developed by digital archivists. We need more cooperation. Where do you start? NARA (National Archives and Records Administration⁶⁹) is one way. A major question before is this: will only the biggest institutions be able to deal with digital preservation? For the smaller institutions, the processes developed by larger institutions are simply not manageable.

DIMAG was developed for archiving digital objects, and BOA for web archiving. Collaboration reduced the complexity by sharing the risks with other memory institutions, on the basis of agreeing to a common object type. But is the agreed-upon standard ISO 15489⁷⁰) actually simpler?⁷¹

In order to provide continual access for content, we need clear specification of data formats. A data format defines an information representation, mapping (bijective) bitstream to contained information. The mapping must be computable. A model for arbitrary data formats inherits the Halting Problem. We have developed a bitstream segmentation graph, which covers 6 types of streams: generic, primitive, structure, transcode, fragment and composite. Each of these form nodes, and there are 3 kinds of edges: segmentation, back transformation, and concatenation.⁷²

⁶⁵ Yan Han*, Chi Pak Chan (University of Arizona). “Modeling Reliability for Digital Preservation Systems.” iPRES 2008, London, 30 September 2008.

⁶⁶ <http://www.dcc.ac.uk/tools/nestor/>

⁶⁷ <http://public.ccsds.org/publications/archive/651x0b1.pdf>

⁶⁸ <http://www.repositoryaudit.eu/>

⁶⁹ <http://www.archives.gov/>

⁷⁰ http://www.arma.org/standards/ISO15489_Pt1.cfm

⁷¹ Christian Keitel* (Staatsarchiv Ludwigsburg). “Ways to Deal with Complexity.” iPRES 2008, London, 30 September 2008.

⁷² Michael Hartle*, Arsene Botchak, Daniel Schumann, Max Mühlhäuser (Technische Universität, Darmstadt). “A Logic-based Approach to the Formal Specification of Data Formats.” iPRES 2008, London, 30 September 2008.

CONCLUSIONS

The first thing we need is a definition of the scope of the impact we wish to have at the University of Alabama, upon policies and procedures for long term access of digital content. The first question is “which content?” If we are only concerned with the digital material being created from Special Collections materials, the problem space is far smaller than if we wish to inform university policies at large. Some institutions are seeking to impact choices made regarding the archival of the university website, management of born-digital content throughout the administration, research and teaching content generated by the faculty, and more. If we want to be consultants to the remainder of the university on how to move forward with regards to long-term access to digital content, we need to educate ourselves and get our own back yard in order.

If we desire a wider impact, I think we need to gain a working understanding of the basic tenets, policies and procedures used by Electronic Document Records Managers, as they have already mapped out the territory as to how to position and frame the effort to gain support and funding from the host institution. Too often people think they are the first ones on the scene, but this is just a new variation (although a complex one!) on a very old theme.

Risk assessment, planning, assessing scope, obtaining buy-in from stakeholders, involving necessary participants, all must take place before any technical implementation on the wider university front. We could certainly position ourselves as leaders, but we need this wider groundwork of involvement and public relations.

Based on the information presented at the conference, I recommend that we investigate potential models for planning and select the one most suitable to our digital services problem space, with an eye towards also identifying models and patterns that may be more suitable for whatever larger problem space we would like to potentially impact. Software is beginning to emerge which may be of assistance in identifying, monitoring, and migrating content; we should test a couple of these as well and determine if they meet our needs. In particular, I recommend that we only consider open source software which is platform-independent, as we may need it to last for a very long time, and may need to alter it to meet our specific needs.

Bit preservation needs to be expanded beyond our current LOCKSS model to incorporate a wider geographic area (preferably different hemispheres) and some variation of platforms. High-end hardware does not appear to be worth the cost. Long term accessibility difficulty can be constrained to the extent that we can normalize the content we seek to preserve to limit the format types to those which are targeted as archival quality. Greater variety in content comes at great risk: while the flash in the pan brings attention and applause, if we cannot provide long term access to the content, we may have wasted our investment for very short term results. It could be very useful to develop an educational component with tools for creating archival-quality copies of research documents at the time of creation.

Development and is becoming more competitive, and best practices are only beginning to emerge. I recommend that we stay informed and involved, so that we can lead our institution and our region toward intelligent choices in the difficult times to come.