

NARRATIVE

National Impact

Libraries have begun to experiment with digital preservation programs. Typically, important paper collections are identified, digitized in uniform format (generally high resolution TIFF files), and stored along with appropriate descriptive, administrative, and structural metadata. Generally the digital materials and metadata are created by experts according to archival specifications and as a result are both high quality and relatively homogenous.

Academic libraries, however, face a far greater range of preservation problems. Image masters may have been created as part of a number of active and inactive digitization projects with different metadata specifications and varying levels of quality control. Students submit electronic theses and dissertations (ETDs) with associated files in an array of media types from spreadsheets to videorecordings. Faculty departments issue newsletters and bulletins in HTML and PDF, and publish research on departmental web servers in Word and LaTeX. The libraries of the state colleges and universities of Florida wish to preserve all of these digital documents and more, with the goal of ensuring their continued usability for as long as possible.

The Florida Center for Library Automation (FCLA), which provides large-scale central data processing systems in support of library operations, proposes to develop a central digital archiving facility for use of the libraries, and indirectly the campuses, of the Florida State Board of Education's Division of Colleges and Universities (DCU).

The FCLA digital archive will utilize current knowledge and best practices in metadata, repository management, and reformatting, to tie these pieces together in an operational facility. This innovative facility would accept submission packages from the state universities, ingest digital documents along with the appropriate metadata, and safely store on-site and off-site copies of the files. Processing would be based upon action plans developed for each physical format. Some objects such as individual TIFF files may be considered "archival" formats and simply preserved until forward migration is necessary. Other objects may have more complex action plans that include the creation of derivative files which preserve the content in more stable formats (called "canonical formats").¹ A primary characteristic of the proposed FCLA digital archive will be its role in serving the real needs of a diverse group of libraries. Every effort will be made to accommodate the formats important to the institutions, whether these are traditionally considered "archival" or not.

A great deal of effort is currently being directed to digital preservation both nationally and internationally. Metadata requirements are edging towards standardization, the OAIS framework is being applied to repositories at leading institutions such as Harvard and Cornell, and a number of ongoing projects are exploring the preservation requirements of specific categories of materials such as government documents and commercial

electronic journals. There are high hopes that the National Digital Information Infrastructure & Preservation Program will build the foundation for a viable national preservation program. However there are few options actually available today to libraries wishing to ensure the long-term preservation of digital materials in their collections, and few (if any) central repositories such as the proposed FCLA digital archive. When operational, the FCLA digital archive will not only provide a valuable service to libraries in the state, but it will also serve as a model for the development of similar facilities nationwide.

Adaptability

Most experts agree that the development of multiple distributed, centralized archiving facilities is the only scalable approach to achieving the level of archiving that will be needed to preserve a significant fraction of important digital content. The FCLA digital archive will instantiate a model that can be replicated by other institutions. The technical and administrative infrastructure developed for the FCLA digital archive will be standards-based to the fullest extent possible, and fully documented for wide dissemination. All software developed by the project will be made available through the World Wide Web for general use. A critical project goal will be the development of realistic cost figures for both short-term startup and long-term management of such a facility. While internally necessary for FCLA to develop cost-recovery pricing, these data will also be a valuable addition to the scant knowledge we have as a profession of the real-world cost of digital archiving.

Design

Preservation strategy

The FCLA digital archive will be based on three preservation strategies:

- 1) preserving the original object for as long as possible;
- 2) creating canonical derivatives for certain formats;
- 3) performing forward migration of both archival masters and canonical derivatives when necessary.

Some data formats (e.g. ASCII, TIFF) can be considered inherently “archival” formats in that the data specifications are well known, the formats are nearly universally supported, and/or a critically large quantity of files exist in the format. There is a reasonable expectation that these formats will continue to be supported for some time, and that when such support is in danger of being discontinued, programs will be available to migrate files in these formats to a current supported version or alternative. Digital objects in such formats will be accepted for archiving with the expectation that forward migration will be performed when necessary.

Other data formats are more ephemeral. Either they change versions relatively frequently (e.g. spreadsheet and word processing formats) or they are specific to some specialty platform or user group (e.g. Macintosh-only formats). The FCLA digital archive may declare some ephemeral formats as ineligible for archiving. However, if possible, the objects will be accepted for archiving, and derivative files will be created that preserve the intellectual content (if not all formatting or functional details) in more stable canonical formats. For example, data and column headers from an Excel spreadsheet may be reformatted into an ASCII file that would retain the meaning of the data even if special Excel features were lost. FCLA will attempt to preserve both the original object and the canonical derivative. Forward migration will be performed on both objects when necessary for as long as it is feasible. The expectation is that at some point use of the original object or its forward-migrated versions may be lost, but the canonical version will continue to be available.

FCLA will maintain an “action plan” for every file format accepted for archiving. Action plans will be sensitive to the version of the format and/or the version of the software used in creating the object in that format. The action plan will indicate a) whether canonical versions will be created upon ingestion, b) the long-term preservation strategy for the original and (if applicable) canonical versions, c) a timetable of anticipated short-term actions, and d) the next date on which the action plan will be reviewed and if necessary revised. For example, the action plan for TIFF images might be a) create no derivative versions, b) migrate to future formats when required, c) take no short term actions, and d) review plan in 3 years. The action plan for an Excel spreadsheet might be a) create a derivative ASCII version on ingestion, b) upgrade the original spreadsheet file whenever a new version of Excel is released c) reformat Excel 2000 documents as Excel 2002 documents in August 2002, d) review plan in 1 year.

Complex digital objects will also be accepted, but must be defined by agreement between FCLA and the depositing library. SGML and XML files will be treated as complex objects consisting of the marked-up text and the applicable DTD or schema. Logical objects in multiple files (e.g. a book digitized as page images) may be treated as complex objects consisting of the set of data files, the structural metadata, and the DTD or schema describing the structural metadata. Regardless of how an object is defined, all components of the object, including any structural metadata required for its use, must be in accepted file formats with action plans.

Functions

FCLA digital archiving services will include the following functions:

1. Contracting with the libraries for service.

In the OAIS model, creators make deposits to archives.² In this application, the libraries of the state universities will make deposits to the FCLA digital archive. The libraries will in some cases be the creators of the digital content that is archived, for example when images created for a library digitization project are archived. In other cases, the libraries

may wish to archive content created externally, for example, student theses. The libraries may archive externally-created content on the condition that copyright for the material is held by the university or some university sub-unit, or if the copyright-holder has explicitly granted the library permission to archive the content. The library will be responsible for any actions (for example, negotiation of service agreements) and fees involved in relation to the content.

Digital archiving services will be provided based on specific agreements between FCLA and each university library, stipulating permissions and obligations on both sides. Agreements can be updated at any time but will be reviewed and re-approved annually.

2. Ingesting objects for archiving.

In the OAIS model, information packages consisting of data objects and associated metadata are the basic units of submission, storage and dissemination. A submission information package (SIP) is submitted to the archive for ingestion, which the archive transforms into an archival information package (AIP) for storage. The archive can then transform the archival information package into a dissemination information package (DIP) for transmission to users external to the archive.

Depositing libraries will submit SIPs to the FCLA archive for ingestion. Currently FCLA uses a local package format for exchanging digital objects with the university libraries, but as part of this project, additional programming will be done to convert from the local format to the Metadata Encoding and Transmission Standard (METS), an emerging standard apparently well-suited for use as an SIP data format.

Following validation by quality control procedures, data files and structural metadata will be stored in an archival directory system, and descriptive, administrative, and technical metadata pertaining to the objects will be extracted and stored in an archival management database.

3. Creation of derivative files in canonical formats.

If an SIP contains one or more objects in formats for which the action plan calls for the creation of canonical derivatives, these will be created upon ingestion and processing of the original. Two AIPs will be deposited in the archival storage system: one containing the original deposited objects, and one substituting the canonical derivatives for the original objects. Complete descriptive, administrative and technical metadata will be maintained for each AIP.

4. Data storage and data management.

Data management will be controlled by an archival management database containing descriptive, administrative and technical metadata pertaining to the archived objects. (Structural metadata is not included here as that is managed as a part of the archived object itself.) Although no single standard for preservation metadata is dominant,

project designers are cognizant of ongoing work in this area and will incorporate appropriate data elements for reference, context and provenance information.³ Technical metadata for image files will conform to the draft NISO *Data Dictionary of Technical Metadata for Still Images*.⁴ The development of appropriate technical metadata specifications for other formats is a major undertaking to which it is hoped this project may contribute.

Back-up copies of archived data will be stored in two physically separate locations. All feasible security protections will be utilized to guard against unauthorized access to or alteration of data. When objects are ingested a checksum will be calculated and stored in the archival management database. At periodic intervals, archived files will be refreshed (physically copied to a new location on media). Before and after refreshment, a checksum for the file will be calculated and compared to the checksum stored in the archival management database. The check before refreshment will confirm that the file has not been deliberately or inadvertently damaged; the check after refreshment will ensure that the file has been successfully copied.

Because this is a “dark” archive, it will be critical to implement mechanisms for ensuring that objects remain usable. General sampling will be done routinely over the universe of all stored objects, in addition to targeted sampling done as part of forward migration.

5. Forward migration.

When a file format is considered to be in danger of obsolescence according to its action plan, all objects in that format will be reformatted if possible to a later version of the format or to some comparable successor format. If this situation arises during the project period, processes will be developed to do this conversion in batch, and samples of the converted output will be tested for usability. When a migration occurs, a new AIP is created and the prior AIP is deleted. Information will be adjusted accordingly in the archival management system

6. Dissemination.

FCLA digital archiving services will not include end-user online access to archived objects for several reasons. First, both security and cost considerations argue that archived objects be stored offline. Second, the objects which are archived will not necessarily be the service versions of digital materials. For example, some digital collections display JPEG and PDF versions of objects, but it is likely that the TIFF images from which these files were derived will be the archival objects. Third, it is anticipated that institutions will want to archive digital objects which are not available to users through any FCLA interface; they may provide their own public interfaces, or the objects may be unavailable to the public.

Copies of archived objects will be supplied to agents of contributing libraries on the written request of the official contact person. For each such request, a mutually

acceptable timeframe for and method of delivery will be negotiated. A METS-based DIP will be used to deliver objects and metadata.

7. Other functions.

Other services that will be provided by the FCLA digital archive will include deaccessioning and both routine and *ad hoc* reporting. In addition, detailed records will be kept of costs incurred for administration, storage and processing (including ingestion, dissemination and reformatting), in order to facilitate the development of a rational charging algorithm that can be implemented at the end of the grant period.

Management plan and Personnel

Because of its mission of providing large-scale centralized data processing services to the libraries of the ten state universities in Florida, FCLA is better positioned than any of the individual libraries to develop a shared digital archiving facility.

The overall project outline calls for year one to focus on the development of a system capable of supporting the first archivable classes of materials, with basic functions of naming, ingestion and data/metadata management. Libraries may begin depositing some classes of objects by month nine, which occurs on a fiscal year boundary. During year two, additional functionality, such as the ability to support additional formats, statistical reporting and cost recording will be added to the service, and the amount of material deposited for archiving is expected to increase. By year three the full functionality of the system and the ability to support most priority materials will be in place, and efforts will be focused on productionizing operations, evaluation, and the development of models for cost-recovery pricing.

The principle investigator and project manager for the project will be Priscilla Caplan, Assistant Director for Digital Library Services, who will devote 25% of her time to the project over the three year period (resume attached). She will coordinate the work of the project team and be responsible for communication between the libraries and FCLA in the development of the archiving service. She will also handle the FCLA side of the negotiation of service agreements with the libraries, oversee technical design and development efforts, manage the evaluation aspects of the program, and work on the development of cost-recovery pricing.

Chris Vicary, Coordinator, Computer Applications, will be the lead programmer/analyst (resume attached). In year 1 he will devote 100% of his time to programming for the digital repository, including migrating the local data exchange format to METS, developing the archival management database, implementing a nameserver system and programming ingest and dissemination functions. In years 2 and 3 he will spend 25% of his time on the project, maintaining the database, developing statistics, productionizing routine functions (such as creating canonical derivatives for certain objects upon ingestion) and responding to “tickler” alerts indicating batches of objects must be

examined or migrated. His work in year 3 will also include taking over any remaining programming for format conversion after the departure of the project programmer.

A full time programmer/analyst will be hired for the project for a 30 month period. The term will begin in month 4 (allowing 3 months for hiring) and end after month 33, allowing the project programmer's functions to be taken over by the lead programmer. The project programmer will be dedicated to the specification and implementation of action plans for various format materials. Implementation of action plans will include identifying and implementing commercial software, or else writing programs locally, for the creation of canonical derivatives and possibly for forward migration. This position will also explore and if feasible implement the creation of digital signatures using public key technology to verify the provenance of archived digital objects.

Craig Lowe, Operations and Quality Control Technician, will devote 10% of his time in year 1, and 50% of his time in years 2 and 3, to the ongoing operation of the digital archive, including running routine and ad hoc statistical reports, performing manual quality control and audit functions, and verifying successful completion of routine functions.

Marty Johnson, Senior Systems Administrator, will devote 5% of his time for the full length of the project to managing file storage for the archive.

Budget and Contributions

See the Detailed Budget, Summary Budget, and Budget Justification included in this application. Please note that the FCLA contribution constitutes 58 percent of direct costs for the project. Also note that commercial software for reformatting is not listed as a funded or matching expenditure; if the decision is made to purchase commercial software, FCLA will cover this expense without reference to the grant.

Project Evaluation

The three primary goals of the project and the evaluation plan for each goal are itemized below.

Goal 1: Establish a working digital archive encompassing all of the functions described in the "Design" section above.

Evaluation of the digital archive will be a major undertaking because there are so many axes of performance to be considered. From the viewpoint of the libraries, administrative functions such as the negotiation of agreements, and technical functions such as ingestion, dissemination, and reporting, must all be working smoothly. It is also important to the library directors that the archive be able to accept the majority (or at least a large proportion) of the types of materials of greatest concern to them. From the viewpoint of FCLA, production processes must be automated as much as possible and record-keeping must be adequate to derive cost figures suitable for use in developing a

realistic charging structure. It will be a major responsibility of the project manager to work with the directors or their representatives in developing a priority list of criteria for functional evaluation and quantifiable indicators of each of these.

However, the most crucial aspect for evaluation is that the libraries of the state university system, as represented by the directors, are satisfied that the archive is trustworthy and reliable. The report “Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources” identifies six attributes of a trusted digital repository⁵:

- Administrative responsibility
- Organizational viability
- Financial sustainability
- Technological suitability
- System security
- Procedural accountability

While the report suggests that a certification program can be developed using this framework, a complete program of this nature does not currently exist. As an interim measure, the Certification (Best Practices) Checklist developed at the Archival Workshop on Ingest, Identification, and Certification Standards will be used to assess the archive both qualitatively and quantitatively.⁶ The checklist will be formally applied in the third quarter of years two and three of the grant period, and results of this assessment will be distributed to the library directors.

Goal 2: Identify costs involved in all aspects of archiving with sufficient granularity to enable the development of reasonable pricing for cost-recovery.

This is extremely complex for a number of reasons. The expense curve for archiving is not constant but peaks at ingestion then flattens at a low level indefinitely. Costs will vary according to format, depending on the frequency of reformatting and storage requirements. Programming and development costs for format conversion must be amortized fairly over an unpredictable number of materials stored in that format.

The mechanisms for measuring costs will be put in place by month 9 of the project. At month 15 (January 2004), the first 6 months of collected data will be used in an exercise simulating the development of a price schedule. Any identified problems with data collection will be corrected at that time, and the exercise will be repeated using 12 months of data in January 2005. This goal will be considered successfully achieved if the second exercise results in an algorithm judged suitable for actual use.

Goal 3: Disseminate tools, procedures and results for the widest national impact.

For the dissemination plan, see Dissemination below. The success of dissemination efforts will be measured by counting external (non-DCU) hits to the project website, counting downloads of code, and counting the number of personal contacts made to project staff requesting information. Because the audience for project information is

limited, 50 external hits per month and 5 or more downloads and information requests will be considered evidence of successful publicity and significant external interest. We expect to reach or exceed this goal in the third year of the project period.

Dissemination

Dissemination is one of the three primary goals of the project, as it is hoped that the FCLA digital archive will serve as a model for the wider development of centralized archiving facilities. A website for the digital archive will be established describing the archive and linking to information for depositors (the DCU libraries) and for the general public. All locally-developed modules will be documented and made freely available for non-commercial use by download from the website. Project reports, documentation of administrative and operational procedures, and cost data will also be made available on the website.

The existence of these tools and reports will be widely announced on relevant lists such as *padiforum-l*, a discussion list for the exchange of news and ideas about digital preservation issues, *lita-l*, the list of the Library and Information Technology Division of ALA, and the JISC *digital-preservation* list. Progress will also be reported to the OCLC/RLG Preservation Commons and to other organizations such as the Coalition for Networked Information (CNI) and the Digital Library Federation (DLF) whose members can be expected to have an active interest in similar issues. The FCLA digital archive has agreed to be a beta test site for Cornell University's PT Watch, should that service be implemented. Towards the close of the project period, an article about the development of the digital archive will be written for submission to D-Lib Magazine (<http://www.dlib.org>).

The budget for the project allows for travel to at least six meetings to exchange information with colleagues active in this area and/or to present reports on the project. It is difficult to anticipate which meetings will be most appropriate given that year three of the project concludes in 2005, especially as workshops on digital preservation are often scheduled as ad hoc (rather than periodic) events. However it is expected that presentations will be given at the fall 2004 CNI Task Force meeting and at the American Library Association meeting in summer 2004. Other likely venues include future DLF Forum and Joint Conference on Digital Libraries (JCDL) meetings.

Sustainability

The start-up costs of the digital archive will be borne during the project period. At the end of the grant period, the outlined functions of the digital archive should be operational, and action plans should be in place for a significant subset of formats the libraries wish to archive. The main programming and analysis activity on an ongoing basis will be to develop action plans for new formats and new types of complex objects as demand arises. This function will be taken over by the lead programmer for the last three months of the project period, and is expected to be sustainable by approximately .25 FTE supported by FCLA. FCLA also plans to provide the permanent staff required to provide continuing

production and operations support for the archive and to carry out administrative functions.

It is the clear intent of FCLA and the DCU libraries to continue the operation of the FCLA digital archive indefinitely. This project should supply sufficient cost data to allow the development and institution of reasonable charging algorithms for cost recovery. This is important not only to guarantee the continued supportability of the archive, but also to rationalize the use of the archive and to provide an economic framework for librarians to make archiving decisions. However, if, for any reason a decision is made not to implement cost recovery pricing immediately at the end of the project period, FCLA is committed to supporting the archive using centrally allocated funding.

REFERENCES

1. Clifford Lynch. "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information." *D-Lib Magazine*, v. 5 no. 9 (September 1999). Available: <http://www.dlib.org/dlib/september99/09lynch.html>.
 2. Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS): Draft Recommendation for Space Data System Standards*. CCSDS 650.0-R-1, Red Book, May 1999. Available: <http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf>
 3. *Preservation Metadata for Digital Objects: A Review of the State of the Art*. A White Paper by the OCLC/RLG Working Group on Preservation Metadata. Available: http://www.oclc.org/digitalpreservation/presmeta_wp.pdf
 4. Data Dictionary: Technical Metadata for Digital Still Images. Working Draft 1.0. Available: <http://www.niso.org/pdfs/DataDict.pdf>
 5. *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources : An RLG-OCLC Report*. Draft for public comment. Mountain View, CA: RLG, 2001. Available: <http://www.rlg.org/longterm/attributes01.pdf>
 6. Archival Workshop on Ingest, Identification, and Certification Standards (AWIICS). *Certification (Best Practices) Checklist*. Available: <http://ssdoo.gsfc.nasa.gov/nost/isoas/awiics/CertifBase.ppt>.
-