

## The Florida Digital Archive and DAITSS: A Working Preservation Repository Based on Format Migration

Priscilla Caplan, Florida Center for Library Automation

This paper was first published as an article in the *International Journal on Digital Libraries*, 20 March 2007 (<http://dx.doi.org/10.1007/s00799-007-0009-6>). The original publication is available at <http://www.springerlink.com>.

**Abstract:** The Florida Digital Archive is a long-term digital preservation repository for the use of the libraries of the public universities of Florida. It is managed by the Florida Center for Library Automation (FCLA) and based on DAITSS, repository software developed by FCLA with the aid of grant funding from the Institute of Museum and Library Services (IMLS). DAITSS is designed to implement active preservation strategies based on format transformations including forward migration, normalization, and localization. As a case study, the Florida Digital Archive and DAITSS show that active preservation strategies can be incorporated into repository applications from the start, and that doing so affects all aspects of application design. The Florida Digital Archive has been in production since November 2005. The DAITSS application is nearing completion and will be released as open source software in 2006.

**Keywords:** DAITSS, Digital archiving, Digital preservation, Florida Digital Archive, Format migration

# The Florida Digital Archive and DAITSS: A Working Preservation Repository Based on Format Migration

## 1. Background

Until recently, higher education in the state of Florida was highly centralized, with ten public universities forming a cohesive state university system under a single Board of Regents. The prevailing spirit was to reduce costs by centralizing administrative functions when possible and limiting local deviations. In that spirit, the Florida Center for Library Automation (FCLA) was established in 1984 to run a single library management system for the use of the ten universities and more generally to centralize support for computer applications in the libraries. Twenty years later FCLA continues to operate a shared library management system, as well a link resolver, a metasearch portal, and other applications. It also provides central hosting services for digital library collections, electronic theses and dissertations (ETDs), and electronic journals.

The libraries of the state university system have always looked to FCLA for digital preservation. In the late 1990s FCLA began hosting collections of books, manuscripts, photographs and other content digitized by the libraries from their Special Collections departments. At the same time it also began storing backup copies of the preservation masters and managing these for fixity and readability. In early 2000 FCLA started planning a more elaborate digital preservation repository service based on the emerging framework for Open Archival Information Systems (OAIS).[1] The immediate impetus for this was the trend in the universities to allow, or in some cases to require, submission of electronic dissertations. The libraries suddenly found themselves responsible for providing perpetual access to born digital resources of great importance to their universities.

In consultation with the library directors, FCLA decided from the start that the preservation repository would be designed for long-term preservation (the initial horizon was 100 years), based on the OAIS reference model [1], and limited to preservation functionality only. Libraries could use any digital library systems, asset management systems, and/or institutional repositories they chose for providing online access to digital content, and they would send only those materials selected for long-term preservation to the FCLA repository. Other early decisions were based on the types of materials the libraries expected to archive. Since most content was expected to be documentary (image, text, audio and video) as opposed to executable (software, games, learning modules), FCLA decided to implement preservation strategies based on reformatting rather than emulation. Since the libraries had little control over the file formats allowed by the graduate schools for ETDs, the repository would have to be able to handle a heterogeneous and expanding list of file formats.

With support from the Institute of Museum and Library Services (IMLS), FCLA designed and developed a preservation repository management system called DAITSS (Dark Archive in the Sunshine State). The repository itself is known as the Florida

Digital Archive (FDA). The FDA went into production in November 2005. As of November 2006, formal agreements have been established with nine of the ten university libraries for the use of the repository, and materials from four of the libraries are being ingested regularly. The majority of these materials have been ETDs and TIFF masters from local digitization projects. Full preservation treatment is available for twelve different file formats: AIFF, AVI, JPEG, JP2, JPX, PDF, plain text, QuickTime, TIFF, WAVE, XML and XML DTD.

## 2. FDA administration and management

Use of the Florida Digital Archive is based on a signed agreement between an institution (called an "affiliate") and FCLA. The agreement lays out the respective responsibilities of each party and clarifies rights in archived materials. The affiliate retains all ownership rights, but must warrant that it is responsible for copyright compliance. The affiliate must have the authority to grant FCLA non-exclusive rights to copy, display and create derivative versions of deposited files. The wording is intended to allow the repository to perform any actions needed to carry out preservation functions, such as viewing or otherwise rendering files when necessary, making backup copies, and creating modified derivative versions.

The requirement that the affiliate can archive only materials for which it can and will grant all rights needed for preservation greatly simplifies the role of repository management and the design of the DAITSS software application. The FDA does not need to encode, store or test permissions, because it can assume that it has all the permissions it needs. It also does not need to manage complex access rights, since the only access to stored materials is via an authorized dissemination request from the institution that deposited the material.

Respective responsibilities are laid out in more detail in the FDA Policy Guide.[2] This describes a model of operation in which responsibility for digital preservation is shared between the FDA and its affiliates. The FDA attempts to ensure that archived materials can be disseminated back to the depositor in renderable form, meaning that the materials can be displayed, played or otherwise used on software currently available at the time of dissemination. The archiving institution is responsible for maintaining its own records of what is archived with the FDA and maintaining adequate descriptive metadata. If the affiliate archives only preservation masters, it is responsible for creating its own service copies and for recreating service copies when the originals become damaged or obsolete.

The libraries of the state universities and certain partners in cooperative digitization projects are eligible to become affiliates. Other organizations and campus units can use the FDA only through an arrangement with an affiliate or with special approval from the policy advisory board. Each affiliate is assigned an account and allowed to name any number of individuals authorized to deposit materials, get reports, and request dissemination and withdrawal of materials. Affiliates also complete and maintain a form

that specifies what materials will be archived for their account and how they should be treated.

FCLA is responsible for the administration and operation of the FDA. The FCLA Advisory Board functions as the policy advisory board for the FDA. The FCLA Board consists of the directors of each of the university libraries, plus the state librarian and a few other *ex officio* positions. As issues of policy arise they are taken to the quarterly meetings of the Board for their recommendation.

The FDA's policies, procedures and documentation have been reviewed internally against the criteria described in the NARA/RLG Audit Checklist for Certifying Digital Repositories.[3] It is the goal of the FDA to become certified as a trusted digital repository when a mechanism becomes available for that process.

### 3. Information packages in DAITSS

Three types of information packages are defined in OAIS: the Submission Information Package (SIP), Archival Information Package (AIP), and Dissemination Information Package (DIP). These correspond to what is submitted to the archive, what is stored in the archive, and what is disseminated from the archive. In DAITSS, every Information Package regardless of type must contain a descriptor and at least one other file. The descriptor is an XML file conforming to the Metadata Encoding and Transmission Standard (METS) that, at a minimum, references each file in the package.[4]

DAITSS requires each intellectual entity (digital book, photograph, dissertation, etc.) archived to be packaged in a well-formed SIP according to a published SIP specification.[5] A single SIP should contain all the files needed to represent a single intellectual entity, and must include exactly one SIP Descriptor. The original SIP descriptor is processed as a descriptor and also archived as a file.

The DAITSS AIP consists of the files archived from the original SIP (including the SIP descriptor), any files created by DAITSS via localization, migration and/or normalization (see *Preservation strategies* below), and an AIP descriptor. The AIP descriptor references all of the files included in the AIP, and contains detailed preservation metadata, including technical metadata describing each file and bitstream, information about relationships among files, and information about events involving the files. (The preservation metadata is also recorded in a MySQL database, but the database version is not considered part of the AIP.)

The DAITSS DIP contains either one or two versions of the intellectual entity. It always contains the original version, that is, the files submitted in the original SIP (including the original SIP descriptor). If one or more files in the AIP has been localized or migrated, the DIP will also contain a complete version of the intellectual entity in which original files are replaced with their most current equivalents. Each version is described in a separate structural map in the METS DIP descriptor.

#### 4. Preservation strategies

The FDA offers two levels of preservation, "full" and "bit." Bit level preservation ensures only that files are preserved intact and are readable from media. Full preservation consists of bit level preservation as well as actions to ensure that the intellectual entity remains renderable as original formats become obsolete: format migration, localization and normalization.

Format migration is the process of creating a version of a file in a more current format, particularly if the format of the source file is in danger of becoming obsolete. Ideally, migration should be as lossless as possible and retain the content, appearance, and behaviors of the source. In DAITSS, migration is performed at the time a file is ingested if a migration routine exists for that file format. It is always possible, however, that a migration routine will become available for a format long after any particular file is ingested. Therefore, as part of the Dissemination process, the Information Package is first routed to Ingest to be re-ingested, at which time any necessary migrations will be performed. To do a mass migration, repository staff would use the reporting system to get a list of all files in the format in question, disseminate all packages containing one or more files in that format, and reingest the packages.

Note that although all preservation treatments take place at the level of the individual file, a file can be accessed only as part of an Information Package -- an individual file can not be disseminated. If a particular file must be migrated, the entire AIP containing the file must be disseminated and re-ingested.

Localization is a term coined by the FDA for the process of creating a version of a file in which all external references have been replaced by relative pathnames. The reason for localization is to ensure that all supplementary files necessary to use or validate a deposited file are available in the repository. For example, if an XML file references an externally stored DTD, an attempt will be made to download that DTD, and to replace the external reference with an internal one. Not all files containing links are localized. For example, a PDF thesis that links to other works in the bibliography will not be localized, because the repository is unlikely to have permission to copy and preserve the linked-to content.

Normalization is the process of creating a derivative of a source file in a format considered to be more "archivable." Many repositories normalize content before ingesting it or as part of the ingest process. For example, the Portico electronic archiving service normalizes incoming source files from publishers to conform to a standard XML DTD.[6] The National Archives of Australia normalizes source objects to create preservation masters in an XML-based archival format.[7] In these and most other examples, materials are normalized to a common, relatively preservable format, in order to minimize the number of migrations needed in the future and to simplify migrations when they must occur. The FDA takes a different approach. If a good normalization

path exists for a particular file format, the FDA will create a normalized version of the incoming file but will not store it. Although this sounds like a waste of processing resources, this method ensures that normalization can be done and catches normalization errors immediately, while sparing the repository the need to store, manage and migrate normalized versions. If for some reason it were to become necessary to use a normalized version, it could be created on demand.

An affiliate must designate whether its materials should be accorded full or bit level preservation treatment. Treatment can be specified based on file format, a "project" code associated with the submission, or both. For example, an affiliate could specify that all of its TIFF files receive full preservation, or only TIFF files for project ABC. Even if requested, full preservation can be performed only for file formats that are supported in DAITSS. If a format is not yet supported, the file will be accorded bit level preservation until full preservation is possible.

Regardless of whether or not derivative versions are created, all original source files (that is, files submitted by the affiliate in the SIP) are maintained unaltered in perpetuity. This serves two purposes. First, it allows a verifiable chain of provenance from the original version to the latest version of the file. Second, it allows for the possibility that a particular migration routine might be superseded by a better one. That is, if a migrated version of file A is created in format X, and later it appears that format Y is a better successor, it may be possible to migrate directly from A to Y rather than migrating from X to Y, decreasing the risk of loss.

## 5. DAITSS Functional Entities

DAITSS is designed as an implementation of the major functional entities of the Open Archival Information System (OAIS) framework: Ingest, Data Management, Access, Archival Storage, and Administration. It also has two additional functions: Preprocessing and Withdrawal. The Preprocessing function ensures that the SIPs submitted by affiliates are valid. Depending on the circumstances it may add, delete, reformat, or otherwise alter the content of a SIP according to custom specifications. The output of Preprocessing is considered to be the official SIP, and goes into the input queue for Ingest. The Withdrawal function removes materials from the repository while retaining some historical information.

Ingest has the task of transforming each incoming SIP into an Archival Information Package (AIP) which can be written to Archival Storage. To create the AIP, it processes each file in the SIP in turn. After checking the file for viruses and verifying the message digest, DAITSS identifies the format of the file and assigns to it a preservation level based on the specifications of the affiliate submitting the package. If the file format is supported in DAITSS, the file is parsed and validated against file format specifications. Technical metadata is extracted and stored in a database. If the format requires normalization, a normalized version is created but not stored. If the format requires migration, a migrated version is created and, like any new file, parsed, validated, and

described in the database. If the format requires localization, a localized version is created, parsed, validated and described in the database.

All files are assigned unique identifiers, as is the AIP as a whole. All significant processing steps and their outcomes are recorded in the database as events associated with the file and/or the AIP. Structural and derivative relationships between the file are also recorded. Finally, an AIP descriptor is created and added to the AIP. The descriptor is redundant with the database and contains all known information about the files, relationships, and events, so that if the DAITSS system were to cease to exist, the archived content would be self-defining. The AIP is written to storage, database updates are committed, and an ingest report is emailed to the contact address for the account associated with the package.

The Access entity has two major tasks: accepting user requests and creating DIPs. Access accepts user requests for reports, withdrawal and dissemination, authenticates the requester, and determines whether the requester is authorized to submit the request. If so, Access reformats the requests and forwards them to the appropriate queue for processing. The Dissemination function within the Access entity creates DIPs in response to dissemination requests. Each DIP contains the content of the original SIP as submitted by the affiliate and, if different, the "last, best" version of the content.

The DIP is created by outputting the requested AIP and routing it back to Ingest as a SIP. Standard SIP processing ensures that each file is examined by the most current version of the software. This might cause some files to be localized or migrated. Also, since classes to support new formats are periodically added to DAITSS, it is possible that previously unrecognized file formats will now be identified. Bug-fixes or other improvements may also cause some changes to the AIP. A DIP is then assembled and routed to the requesting affiliate.

Most of the functions assigned to Archival Storage in OAIS are handled by a third-party storage management system. DAITSS makes calls to storage management according to a generic interface. For the software to use a specific storage management system, an implementation of the generic interface for the specific system must be written. This allows new storage systems to be used with DAITSS without requiring changes in components that use those systems. A configuration option tells the software how many copies of each file to write; all copies are considered equal and treated as masterfiles rather than as a master and backups. The FDA uses IBM's Tivoli Storage Management to write three archival copies to tape. The DAITSS application includes routines to ensure that all files that should be in storage are in storage, that files are readable from media, and that files have not been altered since their message digests were recorded.

DAITSS Administration consists of user interfaces to system and account configuration tables, access to logs and error reports, and methods for populating and cleaning up queues. At this time most administrative functions are done using Unix tools such as vi, but graphical interfaces will be provided in the future.

## 6. Format Handling

Handling of each file format implements an action plan that is based upon extensive analysis of the format, which is written up as a background report. Action plans and background reports are published on the Florida Digital Archive website.[8] Each Action Plan includes:

- whether (and if applicable, how) a normalized version will be created,
- the long term preservation strategy,
- short term preservation actions,
- the date that the Action Plan should be reviewed and possibly revised.

Once the Action Plan is reviewed and approved by the development team, code to handle the format is added to the code base. DAITSS has two hierarchies for format-related information: Data File and Bitstream. In the Data File hierarchy there is a subclass for each type of file such as TIFF, PDF, and JPEG2000. In the Bitstream hierarchy there are subclasses for Audio, Image, Text, Video, Multimedia and Miscellaneous. The work to perform validation, extraction of technical metadata, creation of derivative versions, and other format-specific processing is divided between the Data File classes and the Bitstream classes as appropriate. For example, the Data File class *Jpeg2000* understands and processes the JP2 format as a wrapper of bitstreams, while the Bitstream class *bs.image.jp2image* processes each image inside the wrapper.

The extraction and storage of format-specific metadata is an area that requires judgment because there are no formal or even *de facto* standards for technical metadata for most file formats. As part of the format analysis, the elements of technical metadata that can be extracted automatically from the file or bitstream are identified and reviewed. If it appears the element would be useful in identifying files for selection, or for preserving them through transformations in the future, the element is added to the database schema of the MySQL database and to the XML schema for the AIP, DIP and SIP descriptors. Adding a format requires some expertise and is time and labor-intensive. Ideally, the work to add formats to DAITSS could be shared among a wider group of DAITSS developers. In order to do this efficiently, however, format processing should be made more self-contained and "plug-able" in the code. This change is planned for the next major version of DAITSS.

Designing migration and normalization routines also requires some difficult decisions, particularly about when to use commercial tools, open source tools, or locally developed tools. Use of open source third party tools is preferred, but not always possible. It is often the case that open source tools lag behind their commercial equivalents in functionality, so the decision must be made whether to wait some number of months in the hope an open source tool will be improved, or to go forward with the commercial product.

## 7. Experiences from the first year

At the time of this writing, the Florida Digital Archive has been in production for about a year, from November 2005 through November 2006.

During the first three months of operation, nearly all problems concerned Archival Storage. As mentioned above, the FDA is configured to write three copies of each file in the AIP to tape. Two copies are written locally to a robotic tape unit, and one copy is written in real time over the Internet to a similar tape unit in Tallahassee, about 130 miles away. The software is written in such a way that all three writes must complete before processing can continue.

When the FDA went into production, Ingest immediately came to a stand-still waiting for local tape drives. Although there were four drives in the pool, the archive needed two drives to write the two local copies, and it was contending with database backups, Tivoli data management and the DAITSS system itself attempting to read input data from tape for Prep. This problem was solved by moving to faster drives, whereupon the bottleneck moved to Tallahassee, where contention for the two available remote drives caused the system to time out. To ameliorate this a third remote drive was purchased. After that timeouts were less frequent, but the movement of data was far too slow, and only a handful of packages could be processed in a day. This turned out to be a mismatch between the full duplex setting of the FDA server's network card and the half-duplex setting on the remote switch. With that resolved, the FDA began writing to tape at a steady flow. Finally, in August 2006 we moved onto Florida Lambda Rail which provides a gigabit link between Gainesville and Tallahassee.[FLR] Now the speed of the network exceeds the speed of the tape drives themselves.

With I/O problems out of the way, inefficiencies within the DAITSS code itself became visible. For example, scanning files for viruses takes a relatively long time to complete. The way the application was originally written, every file created by DAITSS (for example, a localized or normalized version) was subject to the all same processing as a file submitted in the original SIP, including virus scanning and other unnecessary operations. That portion of Ingest was recoded to perform only necessary processing on repository-created files.

In the first three or four months of production, the programmers on the DAITSS development team handled all operations as well as problem resolution. Bugs and inefficiencies required recoding and in some cases, redesign. New DAITSS development was put on hold. By month four, FDA operations were transferred to a full time staff member who ensures the system is always running and that data is always flowing through. She monitors production logs for bottlenecks and I/O problems, tracks SIPs through Prep and Ingest, communicates with affiliates, and handles rejected SIPs.

Rejected SIPs are an ongoing operational issue. When a severe error is encountered, SIP processing stops, the SIP is copied to a "rejects" directory, and the submitting account receives an error report. A SIP can be rejected for any number of reasons, but PDF files that cannot be normalized are a recurring problem. DAITSS uses Ghostscript to normalize PDF files into page image TIFFs, and normalization will fail if the PDF uses a

font that Ghostscript does not have. The font has to be located and compiled into Ghostscript before the SIP can be reprocessed.

Despite these problems, Ingest processing runs around the clock, and materials are flowing into the FDA. In the first year of operation (to November 1, 2006) the archive ingested 23,494 SIPs creating the same number of AIPs comprising 190,346 files. Storage requirements totaled 3.7 TB for each set of AIPs; since the FDA writes three masters in two locations, total tape storage used was 11.1 TB.

DAITSS developers, however, are looking forward to doing a major revision of the system for version 2.0. Several planned enhancements are structural, and have to do with simplifying overly-complicated code. Foremost among these is restructuring format processing to make it easier and cleaner to add support for new formats. Other planned developments add functionality, such as support for digital signatures on SIPs, DIPs, and service requests from affiliates. Full support for the PREMIS Data Dictionary, including the ability to export PREMIS metadata in the METS DIP descriptor, is also a priority.[8]

## 8. Conclusion

To date, most digital preservation repositories have focused on bit-level preservation -- gathering content through harvesting or deposit, and managing content to ensure fixity, viability and data integrity. Many repositories intend to add support for active preservation strategies based on migration or emulation sometime in the future.

The DAITSS software used by the Florida Digital Archive was intended from its inception to carry out active preservation strategies based on format transformations: normalization, migration, and localization. The effect of this on the design of the system is pervasive, and goes far beyond the methods that actually create derivative versions of files.

A good example of this is the tracking of relationships. DAITSS assumes the object of preservation is a renderable intellectual entity, and that it does no good to preserve the individual files of a complex object if the larger object itself cannot be rendered. DAITSS treats the content of each SIP as a representation, in the sense defined in PREMIS, as the set of files and structural metadata needed to provide a complete and reasonable rendering of an intellectual entity. The SIP for a book, for example, may contain hundreds of page-image TIFF files and descriptive and structural metadata stored as XML files. The images and the XML must both be preserved for the book to be rendered by page viewing software. Similarly, the SIP for a thesis may contain a PDF file and several linked QuickTime movies. As the SIP is converted into an AIP, relationships among these files must be recorded so that the representation can be reconstructed. If a migrated version of one or more files in the AIP is created, a new representation is also created. The derivative relationship between the source file(s) and the migrated version must be tracked, and so must the relationship among all the files in the new representation. DAITSS Ingest function records these relationships in both the

management database and in the AIP descriptor, and the Dissemination function is able to use them to assemble DIPs containing the appropriate representations.

As of this writing, the Florida Digital Archive has not yet performed a forward migration for any file format. However, normalization is performed routinely on Ingest for several formats. In essence, the only difference between normalization and forward migration is intent. That is, both are format transformations, but the intent of normalization is to create a more "archivable" version and the intent of migration is to create a successor version. Therefore the ability to perform a normalization and to manage the relationships created by that act is a reasonable indication of the ability to perform and manage a migration.

We feel that the Florida Digital Archive moves our understanding of digital preservation forward in two ways. First, it shows that active preservation strategies can be incorporated into repository applications from the start. Second, it shows that doing so affects all aspects of application design.

The DAITSS repository management system will be released as an Open Source Software application in 2006. We hope a larger community of implementers will exercise the software in other settings, and contribute to the code base methods and classes for handling digital formats of importance to them.

#### References:

1. Consultative Committee for Space Data Systems: Reference model for an open archival information system (OAIS). Blue Book (2002), <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>
2. Florida Digital Archive (FDA) Policy Guide. <http://www.fcla.edu/digitalArchive/pdfs/DigitalArchivePolicyGuide.pdf>
3. An Audit Checklist for the Certification of Trusted Digital Repositories. RLG August 2005, <http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf>
4. Metadata Encoding and Transmission Standard (METS), <http://www.loc.gov/standards/mets/>
5. DAITSS METS Document Profile for Submission Information Packages, [http://www.fcla.edu/digitalArchive/pdfs/DAITSS\\_METS\\_SIP\\_Profile.pdf](http://www.fcla.edu/digitalArchive/pdfs/DAITSS_METS_SIP_Profile.pdf)
6. Evan Owens, "A Format-Registry-Based Automated Workflow for the Ingest and Preservation of Electronic Journals," November 8, 2005, Digital Library Federation Fall Forum, Charlottesville, VA. [http://www.portico.org/about/Portico\\_DLF\\_Fall\\_2005.pdf](http://www.portico.org/about/Portico_DLF_Fall_2005.pdf)
7. Florida Digital Archive, <http://www.fcla.edu/digitalArchive/index.html>

7. National Archives of Australia, An Approach to the Preservation of Digital Records. December 2002.

[http://www.naa.gov.au/recordkeeping/er/digital\\_preservation/Green\\_Paper.pdf](http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf)

Florida Lambda Rail, <http://www.flrnet.org/>.

8. Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group. May 2005, <http://www.oclc.org/research/projects/pmwg/premis-final.pdf>