DAITSS, an OAIS-based preservation repository

Priscilla Caplan
Florida Center for Library Automation
5830 NW 39th Avenue
Gainesville, FL 32606
352-392-9020

pcaplan@ufl.edu

ABSTRACT

DAITSS is a preservation repository application developed by the Florida Center for Library Automation (FCLA) for use by the Florida Digital Archive (FDA), a digital repository shared by the eleven universities in the Florida public university system. DAITSS is strictly modeled on the Reference Model for an Open Archival Information System (OAIS). DAITSS can accept a Submission Information Package (SIP), transform the SIP into a stored Archival Information Package (AIP), and transform the AIP into a Dissemination Information Package (DIP) on request. To do so, it directly implements four of the six OAIS functional entities: Ingest, Data Management, Archival Storage, and Access. Functions of the remaining two entities, Administration and Preservation Planning, are performed by FDA staff with support from DAITSS reporting and data management functions. DAITSS is unique among repository applications in that it was designed to ensure the long-term renderability of authentic digital materials. In contrast to Private LOCKSS Networks, which do little but replicate data, DAITSS implements active preservation strategies, maintains standardized preservation metadata including digital provenance, and performs continuous fixity checking on multiple stored copies. The preservation protocol implemented by DAITSS combines bit-level preservation, format normalization, and forward format migration. FCLA will be releasing a second version of the DAITSS application (DAITSS 2) as a sequence of RESTful web services in 2010, at which time the code and documentation will be freely available for other institutions to use.

Categories and Subject Descriptors
J.m [**Computer applications**]: Miscellaneous

General Terms
Design

Keywords
Digital preservation, Preservation repositories, OAIS, DAITSS

1. BACKGROUND
DAITSS is a preservation repository application used by the Florida Digital Archive (FDA), a digital repository shared by the eleven universities in the Florida public university system. The FDA went into production in late 2006, and as of June 2011 it held 290,000 packages comprising 39.1 million files and taking 87 TB storage for a single copy. Four to five terrabytes are ingested monthly.

DAITSS was developed by the Florida Center for Library Automation in Gainesville, an academic infrastructure support organization staffed by professional librarians, programmers, and systems administrators. The directors of the libraries of the state university system were clear that they wanted a preservation solution that would ensure the long-term usability of their non-commercial digital content, including an increasing number of digital dissertations. They were also clear that they did not want to give up their own local content management and "digital library" systems in favor of a shared, central discovery and access system. These requirements mandated a "dark archive" designed for preservation masters of local digital content.

When planning for the FDA began in the early 2000s, there were no vended or open source preservation repository systems available for use, but it was an active time for related developments. MIT's DSpace was released in 2002, offering universities a way to gather local content into their custodial care. The original OCLC Digital Archive was launched in late 2002 supporting ingest on a document-by-document basis. Most importantly, in January 2002 the Consultative Committee on Space Data Systems published the Blue Book *Reference Model for an Open Archival Information System (OAIS)*. In 2003, with support from the Institute of Museum and Library Services (IMLS), FCLA began developing a digital preservation repository application ultimately christened DAITSS, for Dark Archive In The Sunshine State.

2. TECHNOLOGY
The DAITSS application was deliberately designed to meet the requirements for an OAIS as specified in the CCSDS Blue Book, including support for the OAIS information model, functional model, and producer-consumer relationship. DAITSS can accept a Submission Information Package (SIP), transform the SIP into a stored Archival Information Package (AIP), and transform the AIP into a Dissemination Information Package (DIP) on request. To do so, it directly implements four of the six OAIS functional entities: Ingest, Data Management, Archival Storage, and Access. Functions of the remaining two entities, Administration and Preservation Planning, are supported by DAITSS reporting and data management functions.

DAITSS extends OAIS to include the actual implementation of the active preservation strategies resulting from Preservation Planning. The preservation protocol implemented in DAITSS is based on three strategies:

- bit level preservation for the content files contained in the SIP (that is, a copy is retained on readable media without modification, verified by ongoing fixity checking);
- the creation of a normalized copy of source files, when possible, if the files are in formats considered high risk and a good normalization path exists;
- the creation of a migrated copy of files, when possible, if the files are in danger of obsolescence and a good migration path exists.

Normalization is only performed on the original files as contained in the SIP, although a source file may be normalized more than once if better normalization paths are implemented. Migration is performed only on the most current version of a file, so a content file can be migrated successively from format A to format B, and from format B to format C. Intermediate formats created by DAITSS (in this case, the file in format B) are discarded.

Determining what actions to take on files of a particular format requires a great deal of analysis. FCLA employs a formats specialist who studies the file format specification and gathers information about the history and use of the format in order to make informed decisions about what metadata to record and whether derivative versions should be created. These decisions are recorded in "action plans," which are XML documents that can be translated for display or used directly by DAITSS to guide its format-specific processing. If a normalized version is to be created, for example, the action plan will indicate the name of the program to be used and the parameters to send it.

All format transformations are done as part of what DAITSS calls "per file processing." When a SIP is ingested, each file in the SIP undergoes per file processing during which the file is identified and described, and if necessary used as the source of migrated and/or normalized versions. Per file processing is also a step in the Dissemination function, which guarantees that the disseminated package is as up-to-date as possible, and in the Refresh function, which updates packages without disseminating them. Dissemination and Refresh, therefore, effectively implement migration on request and mass migration respectively.

When the Florida Digital Archive began running DAITSS in production, it was the first preservation repository in the United States to implement active preservation strategies based on format transformation, and DAITSS was recognized as a major innovation. What is surprising is that five years later it is equally unique. New technologies have been developed and other large-scale preservation initiatives have arisen, some with the support of the National Digital Information Infrastructure and Preservation Program (NDIIPP). Most of these, however, are geared at gaining physical control of resources for bit-wise preservation. The FDA and DAITSS are nearly alone in implementing active, full preservation in the U.S., where the lion's share of public funding has gone into Private LOCKSS Networks and other storage-based approaches.

This has not been the case in other countries. Findings from a 2009 meeting of representatives from eighteen of the largest providers of digital library systems and services showed a clear differentiation in attitudes towards digital preservation by geographic region. "In North America the practice of digital preservation appears to emphasize long-term storage. Elsewhere (Europe, Middle East, Australasia) digital preservation emphasizes long-term accessibility, readability and understanding."[1] European interest in DAITSS has been high, and FCLA developers closely monitor the PLANETS framework and tool set.

3. ARCHITECTURE AND WORKFLOW

In 2009/2010 DAITSS was completely rewritten as a series of RESTful web services coded in Ruby. The new version, imaginatively called DAITSS 2, went into production in April 2010. This paper describes DAITSS 2, which differs in architecture but not in functionality from the original DAITSS. A diagram of the DAITSS 2 Ingest process is shown in Figure 1.

The function of Ingest is to accept a SIP and convert it into an AIP. A DAITSS SIP must contain the filestreams to be archived as well as a SIP Descriptor describing the contents of the package. The SIP Descriptor is a METS file with certain requirements documented as a METS Profile. All SIPs submitted to DAITSS must come through the Submission Service, which authenticates and authorizes the submitter and validates the SIP to ensure it is well-formed and valid. Checksums provided in the SIP Descriptor are verified. If errors are found the package is rejected. Otherwise the incoming package is assigned an ID number and placed in a workspace.

The workspace is monitored by a program called the Boss, which will initiate an Ingest process when it finds a new SIP, now called a WIP (Workspace Information Package). The address of the WIP is passed to the Ingest handler, which controls its progress through a series of Web services and updates package data in the WIP with information provided by each service. Any problems found from this point on will cause the WIP to be "snafued," stalled in the workspace until an operator takes some action.

The Virus Check Service checks all files in the package for viruses. Following Virus Check, per file processing repeats the services from the Description Service to the XML Resolution Service for every file in the package including the SIP descriptor.

The Description Service incorporates DROID, JHOVE and other tools to identify the format(s) of the file, validate the file against the format specification, and characterize the file using the appropriate technical metadata schema.

The Action Plan Service sends the file to any service required by its action plan, which is a roadmap for processing encoded as an XML document. Files requiring normalization and/or migration

are sent to the Transformation Service which will create a derivative version. XML files will be sent to the XML Resolution Service, which identifies and downloads any external schema needed to validate the XML. New files created by Transformation or obtained by XML Resolution are added to the workspace.

When all per file processing is completed, the Ingest handler builds an AIP Descriptor from the SIP Descriptor and any information added by Virus Check and the per file services. It bundles the descriptors and content files into a tarball for the Storage Service, and also puts a copy of the AIP Descriptor alone into a database called the "XML store".

The Storage Service is responsible for selecting and writing to the appropriate silos (pre-defined sets of storage devices) for long-term storage. A single silo must be all disk or all tape, but DAITSS operators can specify which silos to use in any configuration. The FDA has used all tape, all disk, and a combination of tape and disk silos at different points in time. The number of copies to write is also configurable by the installation. The FDA has DAITSS write two masters on different silos in different geographical areas (Gainesville and Tallahassee). The master copies are then backed up to tape by Tivoli Storage Manager outside of the DAITSS application.
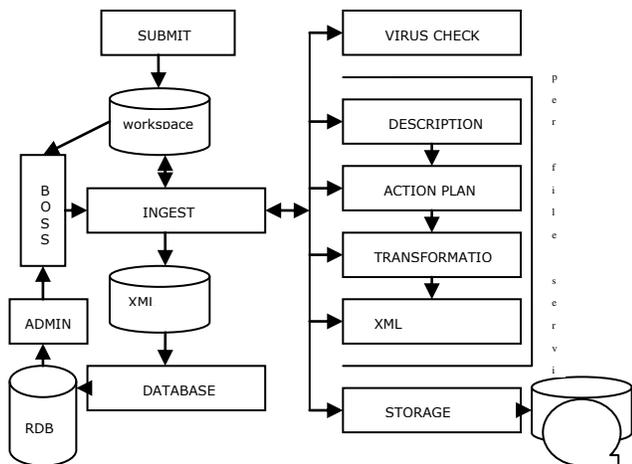

Figure 1. DAITSS 2 Ingest

At this point the Ingest handler is done with the package and removes the WIP from the workarea. The DAITSS Database program, however, monitors the XML store asynchronously for new AIPs. The AIP Descriptor will be parsed and information useful for processing and/or reporting is copied into a relational database for fast access. DAITSS 2 is agnostic to the database management system and has been tested with MySQL and PostgreSQL.

Other processes use many of the same services. Dissemination, Withdrawal and Peek are processes initiated by user requests that

are received by the Request Service and placed in the workspace. Dissemination gets a copy of the stored AIP from the Storage Service and refreshes the contents by repeating the per file processing for each file and updating the AIP. It then creates a DIP from the updated package and sends the updated package to the Storage Service for writing to the silos. The DIP itself is placed in an external location specified by the requestor. Withdrawal removes an AIP from storage but retains historical information in the XML Store. Peek is planned for future DAITSS 2 development. It will retrieve a copy of a package from storage and show it to the requestor without refreshing the contents, a useful function in audit situations. Peek may also be the basis for future cracks of light in the dark repository.

DAITSS 2 also includes functions to support repository management, including configuration, package tracking, and reporting. The application does not include a report writer (the FDA uses Apex and BIRT) but it does maintain tables of operations information as well as fast access tables of AIP information to support reporting. In DAITSS, users were notified by email when their SIPs were ingested or rejected for errors. DAITSS 2 will move from push to pull, and allow authorized individuals to query the system for current status, something long desired by the libraries.

A graphical Dashboard is available to both operations staff and authorized content owners (depositors). The interface allows depositors to submit single SIPs directly to the archive and to list and/or view their own packages, including submitted packages in the process of being archived, rejected packages that have been deleted, and archived packages. FDA staff can also move WIPs in and out of the workspace, release snafued WIPs, and XXXXXX

The FDA and DAITSS are fairly well documented on the FDA website (fclaweb.fcla.edu/FDA_landing_page). In particular, all information about format processing is available, including format action plans and extensive background information, which is important because DAITSS processing is so format-specific. The original DAITSS application was available for downloading under a GNU GPL license, but it was difficult to install and run, and FCLA did not promote the software widely.

DAITSS 2 is more scalable, easier to install, easier to maintain, and much easier to run in a production environment. Current plans are to make DAITSS 2 openly available under an open source license in 2011, but external funding will have to be found for FCLA to provide any significant promotion or support.

4. CONCLUSION

Designing and developing DAITSS and DAITSS 2, and running them in a high volume production operation, has been a tremendous learning experience for all FCLA and library staff involved. The original DAITSS development years were preoccupied with research and modeling of preservation strategies. The first year of production operation required solving a (seemingly) endless chain of bottlenecks impairing throughput. Storage has been a perpetual issue because of cost and the effect of different storage media on all aspects of operations. (For example, the more frequently fixity checks are conducted on

packages stored on tape, the faster the tape will approach its mean time to failure.) As time went on and more content was ingested in production, tracking and reporting increased in importance for both library users and FDA managers and improvements in these areas were required.

The experience of the first five years of running DAITSS indicated the core functionality was satisfactory but the monolithic design of the application was an obstacle to maintaining and testing the code, integrating externally developed applications, and repurposing functionality. These lessons shaped the architecture of DAITSS 2, which has already proven itself to be faster, more reliable, and more maintainable than DAITSS.

The largest lesson learned, however, is simply that it is possible to implement a preservation repository that includes active preservation strategies based on format. It is neither prohibitively difficult nor prohibitively expensive, nor does it require yet-to-be developed technologies. FDA production supporting the eleven universities is handled by one professional librarian and a technical clerical. Operating (as opposed to developing) DAITSS would probably demand part of one programmer and most of a Unix sysadmin; developing actionable action plans for new formats would take another programmer/analyst. An organization running on less of a shoe string would certainly benefit from a position dedicated to training, documentation and user support. These staff resources may be beyond the means of all but the largest cultural heritage institutions but are certainly of reasonable scope for statewide systems and other consortia. The Florida experience stands as a model we invite other organizations to explore.

5. REFERENCES

[1] Beagrie, N. 2003. National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and of Related International Activity.