DigitalPreservationToolsAndStrategies

Contents

Project Description

The goals of the DSpace@Cambridge Digital Preservation Tools and Strategies Project were to create and integrate tools for improving the digital preservation functionality of DSpace, and also to engage in open-ended research that would benefit the wider preservation community.

As of August 30, 2006, the project has been wrapped up. The project, which was originally funded for one year and then extended for another year and a half, was a joint undertaking of the Cambridge University Library and MIT Libraries, in conjunction with the Cambridge University Computing Service.

Among the acccomplishments of the project are the following:

Software Tools

Checksum Checker

A tool for verifying the integrity of bitstreams in the asset store, developed in conjunction with the University of Rochester. This tool is now part of the DSpace 1.4 codebase.

TechMDExtractor

A tool for validating the formats of stored bitstreams, and optionally, for extracting technical metadata from the bitstreams. Harvard University's JHOVE provides the underlying functionality. This tool is awaiting integration into the DSpace codebase, but you can view the documentation TechMDExtractor.

Workflow Pre-ingest Step

An optional workflow step in DSpace that will validate the format of every bitstream upon ingest, and provide the system administrator with extracted metadata for files that are either invalid or not well-formed. Currently JHOVE provides all underlying functionality, although I'm hopeful that someone will expand this step to provide additional functionality, such as virus-checking and migration-on-ingest. This is also awaiting integration into the DSpace codebase; documentation is PreIngest.

Documents

General Documents

- IST_final.pdf Exploring Strategies for Digital Preservation for DSpace@Cambridge, in *Proceedings of the Archiving 2005 Conference*, Society for Imaging Science and Technology, April 2005.
- JhoveLNZComp Why We Chose JHOVE

Format Background Reports

As part of the preservation planning process, we compiled background documents on several formats. These formats are ones we are very interested in preserving, but perhaps just as importantly, they are formats for which there does not yet seem to be a vast amount of information available in the digital preservation community. The background documents depend heavily on foundations laid by three sources:

- the Global Digital Format Registry (GDFR)
- the Florida Center for Library Automation's DAITSS project
- the Library of Congress' Digital Formats pages

The documents are partly an experiment in how to document and categorize format information. The first part of each document consists of data entered into the GDFR prototype Fred (a Format Registry Demonstration), and which therefore conforms to the GDFR data model. The second part consists of data suggested by both the DAITSS project and the LOC project. In some places I've tweaked the headings/categories used by the latter two projects to try to get them closer to something that could fit into a data model, as opposed to a text model, for format information (for instance, the LOC's "transparency" category seems a little broad--it can include statements on encoding/human-readability, and encryption).

- backgrd-HTML.pdf HTML 4.01
- backgrd-XHTML.pdf XHTML 1.0
- backgrd-XLS.pdf MS Excel 10.0
- backgrd-PPT.pdf MS PowerPoint 10.0
- backgrd-MSWord.pdf MS Word 10.0

Preservation Options Reports

These reports summarize the pros and cons of various migration options.

- presOps-HTML.pdf HTML 4
- PresOps-excel.pdf MS Excel 10.0
- PresOps-word.pdf MS Word 10.0