PREMIS in Thought: Transferring preservation metadata

Ardys Kozbial, UCSD Libraries
DLF, April 29, 2008

Credit where credit is due

- UCSD Libraries (MASU)
  - Arwen Hutt
  - Brad Westbrook
- San Diego Supercomputer Center
  - Don Sutton
  - Robert McDonald
  - David Minor
  - Reagan Moore

Context: UCSD Libraries and SDSC

- Collaborative work in digital preservation
  - Long term preservation of video content (NDIIPP/DigArch)
  - LC Pilot Project
  - NDIIPP / Chronopolis

  - http://dpi.sdsc.edu
  - http://chronopolis.sdsc.edu

Context: LC Pilot Project (2006-2007)

- National Digital Information Infrastructure Preservation Program (NDIIPP)
  - www.digitalpreservation.gov
  - Project report:
    - http://www.digitalpreservation.gov/library/reports.html
- Scenario
  - LC is looking for a trustworthy digital repository to manage its assets. Is SDSC that trustworthy repository?
- Building trust
  - Deliverables and tests specified by LC
- From the UCSD Libraries
  - Ardys Kozbial, Arwen Hutt

Parameters for the LC Pilot Project

- Trusted Digital Repository Checklist
  - A1.2 Repository has an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.
    - www.crl.edu
    - Preservation → Digital Archives →Metrics for . . . →TRAC
- Transfer of all deposited data from SDSC to LC
- Transferring preservation responsibility from SDSC to LC
- Protocol must be system neutral, not proprietary
- Assumption: after the data are transferred to LC, SDSC no longer has responsibility for maintenance of them

State Information

- Migration of files
  - Institution name that provided the file
  - Collection name for the record series
  - LC identifier for each file
  - Name used to organize the files at SDSC
  - Physical file name for each file
  - Storage location for each file
  - LC checksum for each file to verify integrity
  - SDSC checksum for each file
  - Date SDSC checksum was validated
  - Status of transfer of file from LC
  - Date file was received at SDSC
  - Number of replicas
  - Location of each replica
  - Creation date for each replica
  - Checksum for each replica
  - Synchronization date for each replica
- Data integrity
  - Logging of all errors for each collection
  - Logging of all errors for each storage system
  - Name of procedure for recovering from each error type
  - Logging of execution of recovery procedures
  - Result of execution of each recovery procedure
  - Validation of consistency of the metadata catalog (file exists for each record)
  - Validation of consistency of the storage vaults (record exists for each file)
  - Dates of consistency checks
  - Most recent date all checksums have been verified
  - Most recent date all replicas have been synchronized
  - Location of metadata catalog backups
  - Most recent date metadata catalog backup created
  - Location of metadata catalog log file

Highlights
File Preservation Transfer Report: Standards

- What information is needed to effectively transfer preservation responsibility for the files themselves?
- Use the data standards supported by LC
  - METS
    - Content packaging standard
    - Does not place restrictions on schemas
    - The METS Profile communicates rules about content and construction of METS objects.
    - METS is used to document this File Preservation Transfer Package
  - PREMIS
    - Use of metadata to support digital preservation
    - Does not proscribe how information is expressed
    - Data dictionary is valuable for identifying existing metadata which satisfies requirements of the standard (SDSC State Information)

Highlights
File Preservation Transfer Report: Scope

- Not relevant
  - Data used to describe the specific repository environment, but that are not intrinsic to the file outside of that repository context.
    - Example: storage location of replicas
- Relevant
  - Preservation processes that were applied to the file

Highlights
File Preservation Transfer Report: Characteristics

- Descriptive metadata
  - None provided in this context, rather, a link to the LC Prints + Photographs database
- Technical and digital provenance metadata
  - Technical characteristics of the file
    - Can be extracted from file headers
  - Preservation events associated with the file
    - Examples: ingestion, fixity check
  - Identification of agent(s) responsible for an event

Questions Outstanding

- Not implemented
- Procedures for handling file versions created as part of the preservation function should be explored.
- Development of controlled value lists for event types, event outcomes, etc. to facilitate consistent application of terminology.
- Although it was developed for all file preservation transfer needs, it was created in the context of a particular scenario – image files. Therefore it needs more testing.

Context: Chronopolis

- National Digital Information Infrastructure Preservation Program (NDIIPP)
  - www.digitalpreservation.gov
  - Project description:
    - http://chronopolis.sdsc.edu
- Scenario
  - Chronopolis is the preservation repository for data. The client wants to get some or all of its data out of the repository.
- Preservation Service Providers
  - SDSC, UCSD Libraries, University of Maryland, NCAR
- Clients
  - CDL (web crawls), ICPSR (social science data)

Process for Chronopolis

- The Preservation Service Providers use SRB as the storage management system.
- Start with metadata that are collected by SRB.
- Figure out which metadata need to follow the data.
- Map these metadata to PREMIS elements.

- Policies.
  - Are all the replicas equal?
  - How are data errors fixed and documented?
  - How are data errors reported to the clients?
  - How is data integrity reported to the clients?

More information

akozbial@ucsd.edu

chronopolis.sdsc.edu