Robust Technologies for Automated Ingestion and Long-Term Preservation of Digital Information

PI: Joseph JaJa

Co-PIs: Allison Druin and Doug Oard

Institute for Advanced Computer Studies

University of Maryland, College Park

Overall Goals

- Development of tools and technologies for automated ingestion and management of preservation processes.
- Evaluation and demonstration of tools on widely different collections.
- Overall architecture based on open standards and web technologies, and will interoperate with data grid and digital library technologies.
- Overall approach captures all essential elements of the Open Archival Information System (OAIS) Reference Framework.

Main Technology Issues

- Management of Technology evolution:
  - Storage, Information Management, Representation, and Access.
- Ensuring long term authenticity of and access to digital assets.
- Risk Management and Disaster Recovery:
  - Technology degradation and failure;
  - Natural disasters such as fires, floods, etc.
  - Human-induced operational or malicious errors.

ADAPT: Approach to Digital Archiving and Preservation Technology

- Digital object model that encapsulates content, structural, descriptive, and preservation metadata.
- Layered software architecture based on three levels of abstraction: data, information, and preservation.
- Organized to enable collaborative, community-based efforts such as replication, “dark archiving”, and Global Digital Format Registry.
- Components expressed within the Open Archival Information System (OAIS) reference framework.

Visual History Foundation Collection

- 52,000 testimonies (116,000 hours of video) in 32 languages from Holocaust survivors, amounting to 180TB of MPEG-1 video.
- Over $16M were invested to create a rich collection of metadata to enable fast content-based access.
- UMD has access to significant portions under an NSF ITR project (Doug Oard)

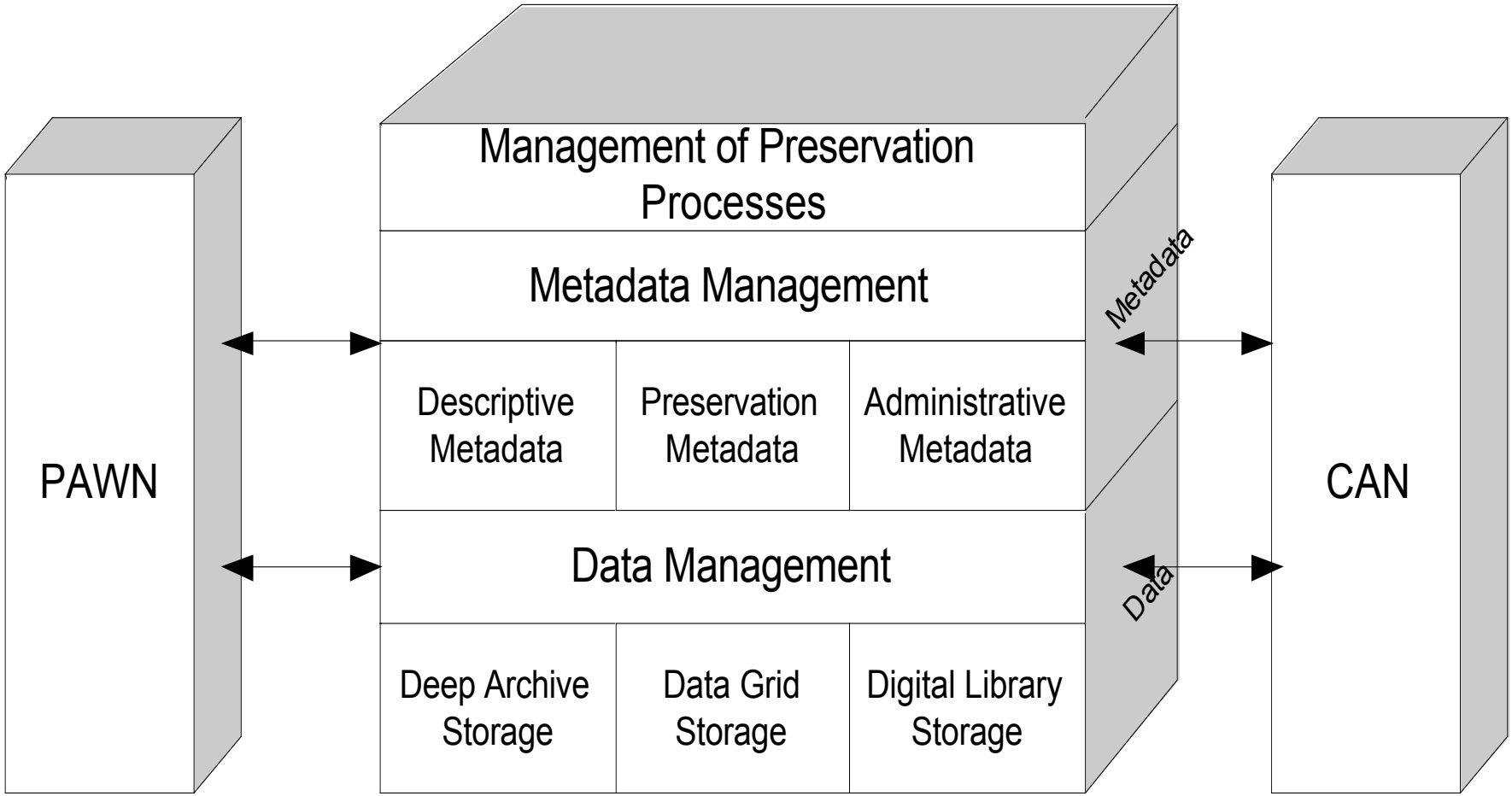
International Children's Digital Library (ICDL)

- Joint project between UMD and the Internet Archive funded by NSF and IMLS (Allison Druin).
- Goal: efficient search, browsing, and reading of a collection of 10,000 books in 100 languages.
- Current holdings over 600 books in 30 languages, with innovative book readers and browsing tools.
- Books are digitized in TIFF format, and processed in 5 sizes of JPEG2000 for each page of each book.

Global Land Cover Facility (GLCF)

- Over 15 TB of Earth Science data organized in a digital library with advanced tools for searching, browsing, and visualization.
- Holdings are raster images or in vector formats using FGDC (Federal Geographic Data Committee)-compliant metadata.
- Multiresolution, multi-temporal, and multispectral imagery with wide varying sizes.

Main Software Components of ADAPT:



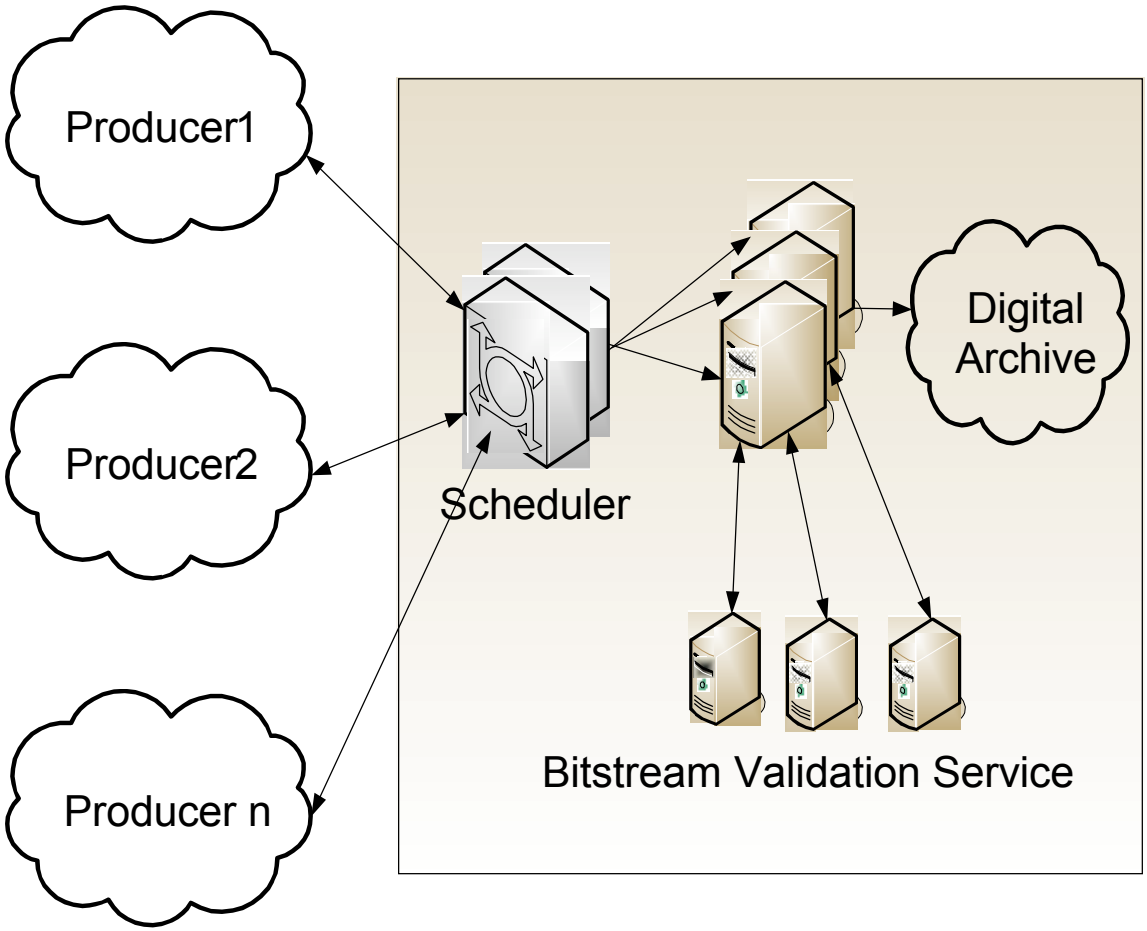
Pilot Persistent Archive Project (SDSC, UMD, and NARA)

- Based on the SRB data grid middleware:
  - Logical name spaces, provide global identifiers for records, users.
  - Data access through storage system abstraction
  - Replication for disaster back up.
  - Uniform access, common user interface across file systems and archives.
  - Single sign-on for users, authentication across administration domains.
- Data grid federation
  - Replicate records and preservation metadata between data grids.

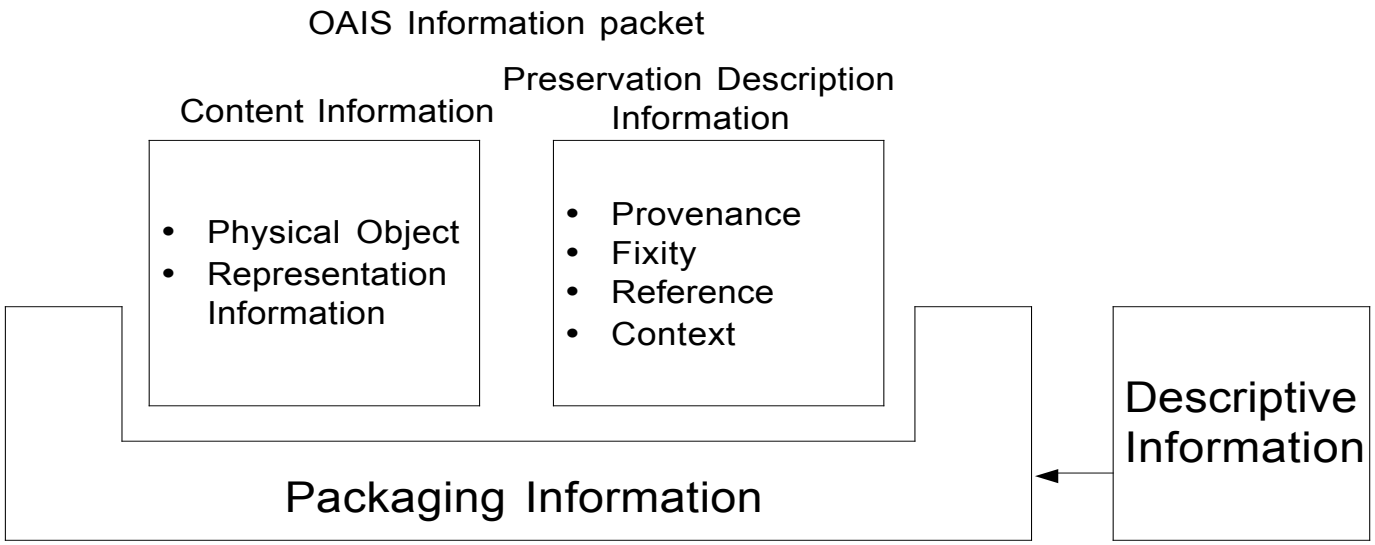
Producer – Archive Workflow Network (PAWN)

- Distributed and secure ingestion of digital objects into the archive.
- Use of web/grid technologies – platform independent
- Ease of integration with data grids or digital libraries.
- XML Representation of metadata and bitstream
  - Self describing bitstream submissions
- Accountability of transfer and guarantee of data integrity

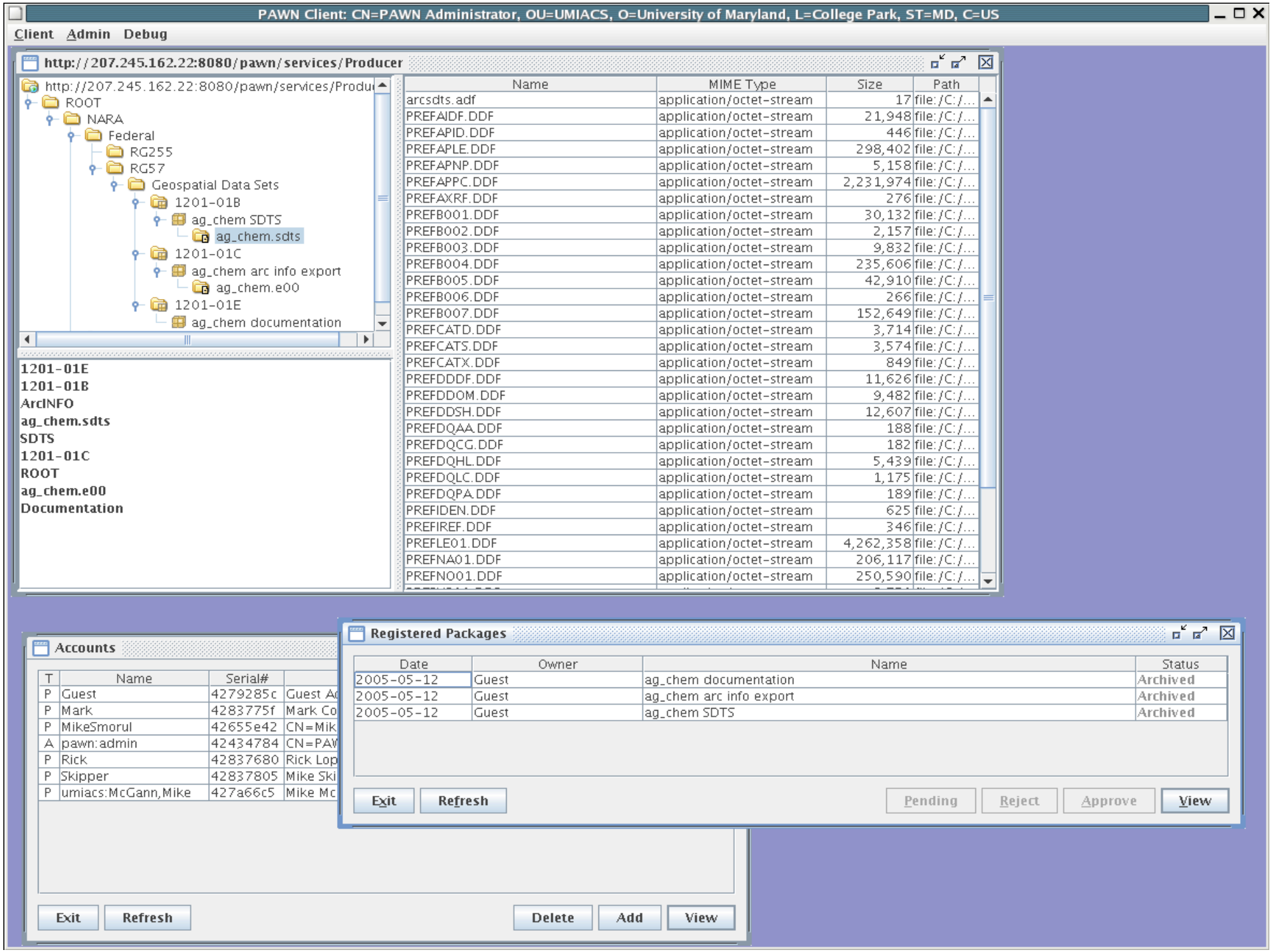
More About PAWN

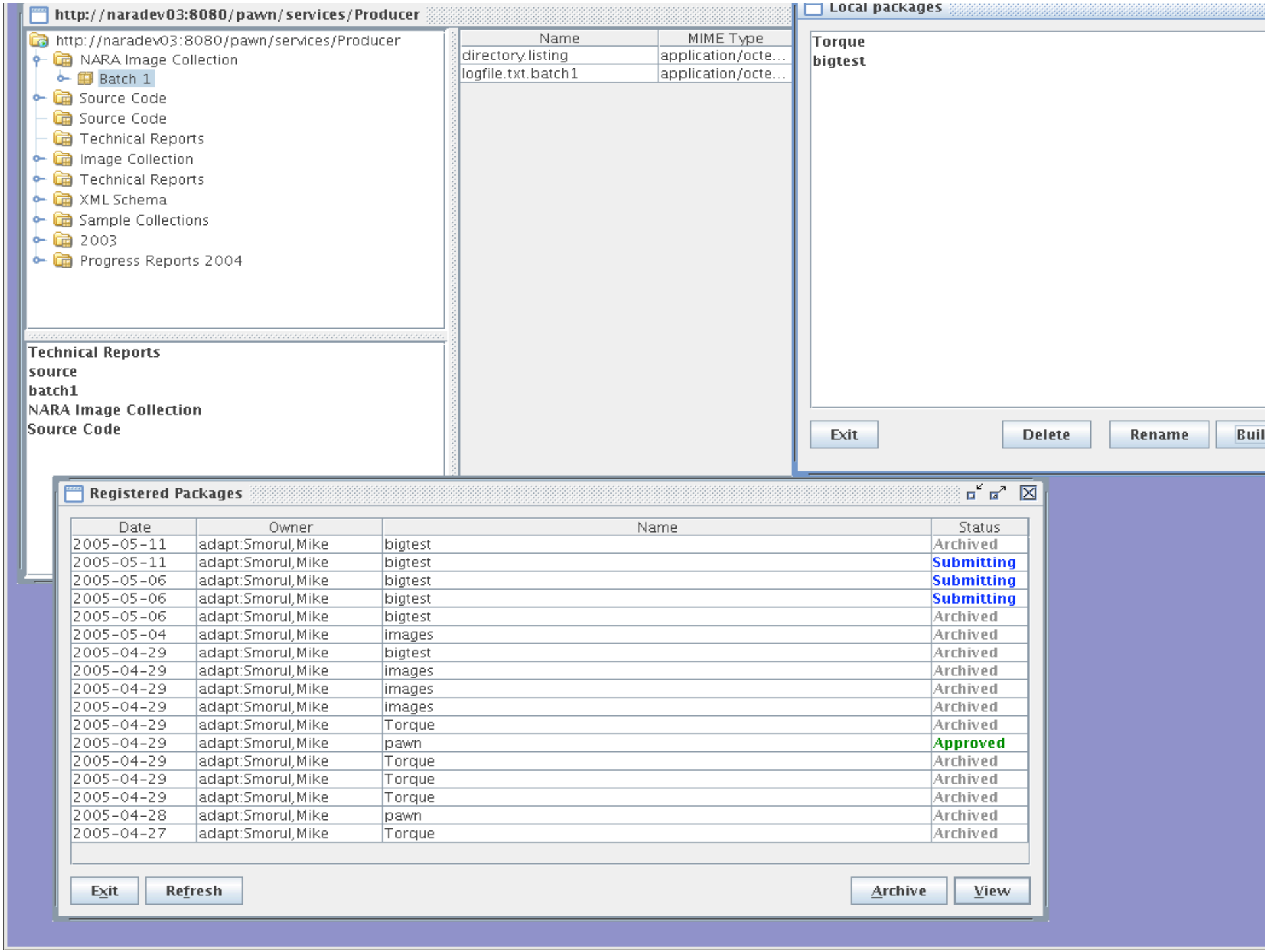


Submission Information Packet (SIP)



- METS Handles all areas of a SIP except Physical Object and Descriptive Information
- Descriptive Information can be embedded into METS as 3rd party XML schema
- Submission agreement constrains how a SIP is structured and described.


PAWN Client: CN=PAWN Administrator, OU=UMIACS, O=University of Maryland, L=College Park, ST=MD, C=US
Client Admin Debug
http://207.245.162.22:8080/pawn/services/Producer
http://207.245.162.22:8080/pawn/services/Produ
ROOT
NARA
Federal
RG255
RG57
Geospatial Data Sets
1201-01B
ag_chem SDTS
ag_chem.sdts
1201-01C
ag_chem arc info export
ag_chem.e00
1201-01E
ag_chem documentation
1201-01E
1201-01B
ArcINFO
ag_chem.sdts
SDTS
1201-01C
ROOT
ag_chem.e00
Documentation
Name MIME Type Size Path
arcsdts.adf application/octet-stream 17 file:/C:/...
PREFAIDF.DDF application/octet-stream 21,948 file:/C:/...
PREFAPID.DDF application/octet-stream 446 file:/C:/...
PREFAPLE.DDF application/octet-stream 298,402 file:/C:/...
PREFAPNP.DDF application/octet-stream 5,158 file:/C:/...
PREFAPPC.DDF application/octet-stream 2,231,974 file:/C:/...
PREFAXRF.DDF application/octet-stream 276 file:/C:/...
PREFB001.DDF application/octet-stream 30,132 file:/C:/...
PREFB002.DDF application/octet-stream 2,157 file:/C:/...
PREFB003.DDF application/octet-stream 9,832 file:/C:/...
PREFB004.DDF application/octet-stream 235,606 file:/C:/...
PREFB005.DDF application/octet-stream 42,910 file:/C:/...
PREFB006.DDF application/octet-stream 266 file:/C:/...
PREFB007.DDF application/octet-stream 152,649 file:/C:/...
PREFCATD.DDF application/octet-stream 3,714 file:/C:/...
PREFCATS.DDF application/octet-stream 3,574 file:/C:/...
PREFCATX.DDF application/octet-stream 849 file:/C:/...
PREFDDDF.DDF application/octet-stream 11,626 file:/C:/...
PREFDDOM.DDF application/octet-stream 9,482 file:/C:/...
PREFDDSH.DDF application/octet-stream 12,607 file:/C:/...
PREFDQAA.DDF application/octet-stream 188 file:/C:/...
PREFDQCG.DDF application/octet-stream 182 file:/C:/...
PREFDQHL.DDF application/octet-stream 5,439 file:/C:/...
PREFDQLC.DDF application/octet-stream 1,175 file:/C:/...
PREFDQPA.DDF application/octet-stream 189 file:/C:/...
PREFIDEN.DDF application/octet-stream 625 file:/C:/...
PREFIREF.DDF application/octet-stream 346 file:/C:/...
PREFLE01.DDF application/octet-stream 4,262,358 file:/C:/...
PREFNA01.DDF application/octet-stream 206,117 file:/C:/...
PREFNO01.DDF application/octet-stream 250,590 file:/C:/...
Accounts
T Name Serial#
P Guest 4279285c Guest Ac
P Mark 4283775f Mark Co
P MikeSmorul 42655e42 CN=Mik
A pawn:admin 42434784 CN=PAW
P Rick 42837680 Rick Lop
P Skipper 42837805 Mike Ski
P umiacs:McGann,Mike 427a66c5 Mike Mc
Exit
Refresh
Delete
Add
View
Registered Packages
Date Owner Name Status
2005-05-12 Guest ag_chem documentation Archived
2005-05-12 Guest ag_chem arc info export Archived
2005-05-12 Guest ag_chem SDTS Archived
Exit
Refresh
Pending
Reject
Approve
View


http://naradev03:8080/pawn/services/Producer
http://naradev03:8080/pawn/services/Producer
NARA Image Collection
Batch 1
Source Code
Source Code
Technical Reports
Image Collection
Technical Reports
XML Schema
Sample Collections
2003
Progress Reports 2004
Name
MIME Type
directory.listing
application/octe...
logfile.txt.batch1
application/octe...
Technical Reports
source
batch1
NARA Image Collection
Source Code
Local packages
Torque
bigtest
Exit
Delete
Rename
Registered Packages
Date
Owner
Name
Status
2005-05-11 adapt:Smorul,Mike bigtest Archived
2005-05-11 adapt:Smorul,Mike bigtest Submitting
2005-05-06 adapt:Smorul,Mike bigtest Submitting
2005-05-06 adapt:Smorul,Mike bigtest Submitting
2005-05-06 adapt:Smorul,Mike bigtest Archived
2005-05-04 adapt:Smorul,Mike images Archived
2005-04-29 adapt:Smorul,Mike bigtest Archived
2005-04-29 adapt:Smorul,Mike images Archived
2005-04-29 adapt:Smorul,Mike images Archived
2005-04-29 adapt:Smorul,Mike images Archived
2005-04-29 adapt:Smorul,Mike Torque Archived
2005-04-29 adapt:Smorul,Mike pawn Approved
2005-04-29 adapt:Smorul,Mike Torque Archived
2005-04-29 adapt:Smorul,Mike Torque Archived
2005-04-29 adapt:Smorul,Mike Torque Archived
2005-04-28 adapt:Smorul,Mike pawn Archived
2005-04-27 adapt:Smorul,Mike Torque Archived
Exit
Refresh
Archive
View

Management of Preservation Processes:

- Policy driven management of preservation processes.
- Main Components:
  - System Registry: available data/metadata repositories; supported file formats; certified transformations.
  - Registry of Policies: replication, refreshing, and migration.
  - Monitoring System to evaluate the archive's health on a regular basis.

Deep Archive

- Erasure codes are forward error correction codes that transform an input object into fragments such that only a specific number of arbitrary fragments can be used to reconstruct the object.
- Using a peer to peer DHT scheme, distribute the fragments among the nodes.
- Integrity and survivability of each object is guaranteed with high probability (can also be made unforgeable and self-verifying).

Consumer – Archive Network (CAN):

- Enables long-term access and information discovery across collections.
- Manages retrieval and display of content.
- Leverages advanced digital library services.
- Grid Retrieval and Search Platform (GRASP) prototype.

Evaluation Strategy

- Ingestion tools: based on tests to evaluate usability, functionality, performance, and interoperability.
- Management of perservation processes: tests to determine functionality in setting and controlling system and policy registeries; performance tests for auditing, tracking services, …
- Deep archive: reliability, ability to retrieve data, and performance.

Conclusion

- Focus on tools and technologies for ingestion and management of preservation processes with core functionality.
- Demonstration and evaluation using rich collections of different types of data.
- Interoperability with the SRB middleware and digital library technology such as DSPACE and Fedora.