A DSpace-based Preservation Repository Design

Open Repositories 2007

Joseph G. Pawletko, Ekaterina Pechekhonova, Unni Pillai

New York University Libraries
Presentation Overview:

- Motivation
- Architectural Questions
- Our Approach
  - Background: OAIS Reference Model
  - Implementation details
- Current Status and Future Development
- Q & A
Motivation:

- NYU’s Digital Library Program is working on several digital preservation grants
  - Hemispheric Institute Digital Video Library (HIDVL)
  - Afghanistan Digital Library (ADL)
  - NDIIPP: Preserving Digital Public Television (PTV)

- Grants have similar objectives
  - preserve and provide access to content
Architectural Questions:

- Build separate, grant-specific applications or a single Preservation Repository (PR) to fulfill grant requirements?

- Build monolithic or distributed applications?
Our Decision:

- Build a single **Preservation Repository**
  - distributed architecture
  - loosely coupled **components** (stable interfaces)

- Each **component** provides subset of PR functionality
  - build **project-independent components** if possible
  - build **project-specific components** when required
Why Components?

- Technology is changing rapidly…
- Using components…
  - provides upgrade flexibility
  - improves software development return
  - decreases wasted development effort
OAIS Functional Entities [2]

- Ingest
- Archival Storage
- Access
- Descriptive Info
- Data Management

Common Services

Preservation Planning

Management

queries
result sets
orders

SIP
AIP
AIP
DIP

2007-01-24
J.G. Pawletko - NYU Libraries
NYU-DL PR Implementation

Data Management
- DSpace Features

Archival Storage
- DSpace Item Importer
- DSpace API
- Maintenance API

DSpace

Access: Discovery
- OAI-PMH
- SRW
- SRU

Access: Dissemination
- Project-Specific

Ingest
- Project-Specific

Common Services
- Project-Independent
Ingest, Archival Storage, Data Management

- Input Monitor
- AIP Generator
- QA
- Ingest Manager
- Ingest Prep
- Archive Ingest
- Cross Check
- SIP
- AIP Staging
- Arch. Storage Data Mgmt.
- database entries
- ingest support files

- 2007-01-24
- J.G. Pawletko - NYU Libraries
Large File Ingest

- **PTV** project must ingest large files
  - High Definition source files can run 800GB per hour

- Files of this size cannot be handled by standard DSpace Ingest mechanism

- Using **SRB** features available since DSpace 1.3
  - store content in SRB
  - register content in DSpace

- Have successfully ingested files up to 130GB
AIP Structure / Update / Rollback

item_METS.xml
(item_METS_v001.xml copy)

item_METS_v001.xml

DMD
TMD foo
.foo

item_METS_v002.xml

TMD bar
.bar
Maintenance

Scope:
- enable authorized users to manage repository data and metadata
- allow update operations to be rolled back

Leverage existing components
- DSpace APIs
- Authentication Component
- Authorization Component
The **Maintenance Interface**:

- provides an **archival-storage/data-management independent** Java interface for PR Maintenance Operations

- can be invoked by a **GUI**, a **WebApp**, or through the **Command Line Interface**
NYU-DL PR Common Services

Common Services

**Authentication**
Project-Independent Scripts

**Authorization**
Project-Independent Scripts

**Persistent Identifier Management**
Project-Independent Scripts

**Persistent Identifier Resolution**
CNRI Handle System®*

* HANDLE SYSTEM is a registered trademark of the Corporation for National Research Initiatives
Authentication Component

- Want to **centralize** Authentication functionality
  - implement once for all web applications, not once for each web application

- Built a **Shibboleth**-based Centralized Authentication Component
  - based on **Yale Central Authentication Service**[^3] (CAS) design
  - leverages Database of Recorded American Music (DRAM) Shibboleth work at NYU
Authentication Component

User (via browser)

Web App

user_id

Shibboleth

WAYF IdP SP

authn.cgi

backend server script

Authentication Interaction:

- Web App redirects to Shibboleth-protected script
- User authenticates via Shibboleth infrastructure
- authn script
  - extracts user_id
  - generates ticket
  - stores (ticket, user_id) pair
  - returns ticket to Web App
- Web App exchanges ticket for user_id via XML-RPC
  - (ticket, user_id) pair deleted
- web app uses authenticated user_id in application
Authorization Component

- Built centralized Authorization service

- Interaction:
  - client sends request parameters to Authorization via XML-RPC
  - Authorization component determines:
    - User’s Role in Collection
    - Permitted Actions for Role
    - Permit/Deny verdict
  - Authorization returns verdict

Client Application

Authorization Script

(user_id, action, item, collection)

Permit or Deny

Authorization

User → Collection
→ Role

Role → Permitted Actions
Persistent Identifier (PID) Manager

- Allocates and manages persistent identifiers
  - Handles, Noids, etc.

- Interaction:
  - Client requests an identifier operation via XML-RPC
    - e.g., create, update, etc.

- Persistent Identifier Manager:
  - interacts with identifier-specific applications to satisfy request
  - returns status, data to Client
Access: Discovery

- **OAI-PMH** in standard DSpace distribution (OAICat from OCLC)
- **SRW/U**: DSpace add-on module available from OCLC (Ralph LeVan at OCLC)
Access: SRW/U Extension

- Extended SRW/U to return real-time mapping using SRW/U protocol’s “extra data” fields
  - original filename → bitstream location

- **Disseminators** use mapping to access METS files, extract metadata, and build links to content

- Real-time map eliminates need to update AIP METS files when changing filesystems
Access: Dissemination

User clicks on Handle URL

Handle System™

Resolver

Disseminator

DSpace Asset Store

SRU

AIPs

DIP
PR Status

- Built **Proof-of-Concept PR** from Core Components

- Tested:
  - DSpace Item Importer, SRU Discovery, Handle Resolution, Resolver Functionality, Dissemination, DSpace interaction with External Local Handle Server
PR Status & Future Plans

- Currently building remaining functionality
  - Ingest Scripts
  - Maintenance API
  - PID Manager

- Upgrading components for Production
  - Authorization
  - Authentication
  - SRW/U
  - Resolver
  - Disseminator

- Plan to load pilot data into the full system 1H’07
Acknowledgements:

- Andrew W. Mellon Foundation
- National Endowment for the Humanities
- Library of Congress

- Dean Carol Mandel
  - Dean of New York University Libraries

- Dr. David Ackerman
  - Executive Director .edu Services and Digital Library Program, NYU

- James Bullen
  - Head of the Digital Library Program, NYU

- Sayeed Choudhury, Tim DiLauro
  - Johns Hopkins University

- Sun Microsystems
Wrap up…

Questions?

Thank you for your time!
References


[2] Ibid., page 4-1
