Fedora: Complex Objects, Information Networks, and the Challenges of Digital Preservation

iPRES
October 9, 2006

Sandy Payette
Co-Director, Fedora Project
Researcher, Cornell Information Science
The Fedora Project

Fedora

- Flexible
- Extensible
- Digital
- Object
- Repository
- Architecture

History

- Cornell Research (1997-)
  - DARPA/NSF-funded research
  - Reference implementation
  - Interoperability experiments w/CNRI

- Open Source Project (2002-)
  - Andrew W. Mellon Foundation
  - Educational Community License (ECL)
  - DLF support for Fedora community
  - Joint development
    - Cornell University
    - University of Virginia
Networks of Complex Objects

• Documents
• Text
• Data
• Simulations
• Images
• Video
• Computations
• Automated Analyses
Fedora: Meeting requirements of today's applications

- Institutional repositories
  - ARROW (http://www.arrow.edu.au/)
- E-Science
  - DART (http://dart.edu.au/about/)
- Next-generation scholarly communication
  - eSciDoc (http://www.escidoc-project/de/homepage.html)
- Open Access Publishing
  - PLoS ONE (http://www.plos.org/cms/node/36)
- Collaborative Digital Libraries
  - NSDL 2.0 (http://nsdl.org/)
Fedora
addressing the preservation challenges...
Strategy/Approach

• We provide an open-source repository service that is flexible and preservation-enabling

• We provide an open-source service framework to enable collaborative development of a trusted repository environment

• We provide a core set of value-add services

• Others to contribute new services as needed
Fedora Service Framework

Fedora Services

- PROAI
- OA Provider
- Search
- Directory Ingest
- Fedora Repository Service
- Workflow
- Messaging
- Preservation Monitoring
- Preservation Integrity
- PID Resolution
- Other
- Replication

Apps

- Administrator
- DirIngest
- Others
- Fez, DPubs, others

Specialized Clients

- Enterprise system paradigm
- Web services
- XML
First, the Core Repository Service...
Fedora Core Repository Service
(OAIS Perspective)

Preservation Planning

Core Fedora Repository Service
- Management
- Ingest
- PolicyEnforce
- Validation
- Access
- Dissemination
- ResourceIndex
- Storage
- Registry

Producer
- METS
- FOXML
- Future

Web Service

Consumer
- RDF query
- triples
- metadata query
- result set

Administration

objects (xml)
byte streams
SQL registry + metadata
rdf-based index

Producer
- API-M
- Producer
- Web Service

Consumer
- API-A
- Consumer
- Web Service

getDissemination
DIP

Producer
- Producer
- Web Service

Consumer
- Consumer
- Web Service

Administration

The Fedora Digital Object

- The basic unit of management, *encapsulating all essential information* about the entity

- Can exist independent of the repository operating environment

- Based on standards and non-proprietary formats to minimize dependencies
  - XML
  - RDF
  - XACML (XML-based security policies)
  - URIs
The Basics for Supporting Preservation

- XML-based Ingest and Export (METS, FOXML, other)

- XML-based Digital Object Storage

- Automatic Versioning of content “datastreams”

- Automatic Audit Trail of all modifications to objects

- Integrity Checking (v2.3)
  - On-demand
  - Auto via repository config
What about the Network of Digital Objects?
Object Relationships as RDF/XML

- Use RDF/XML to express relationships
- RDF is naturally extensible via new ontologies
- RDF stored inside digital object XML files (RELS-EXT)

- Graph index updated on ingest/modify/delete
- **Rebuild** by crawling digital object XML files
It’s all in the XML…

- **Repository Rebuild Service**
  - Reconstitutes the repository in case of corruption or disaster
  - How? By crawling persistent XML object store
  - Rebuilds all repository registries (relational db)
  - Rebuilds the graph index (RDF-based triple store)
  - Rebuilds internal search index (relational db)

- **One requirement…**
  - Assumes simple backups of the XML object store
Moving forward:
Fedora Preservation Working Group

Preservation Support Services
Membership in the WG

- Grace Agnew - Rutgers
- Paul Bevan - National Library of Wales
- Dan Davis - Harris Corporation
- Kevin Glick - Yale
- Ron Jantz (chair) - Rutgers
- Sandy Payette - Cornell
- Eliot Wilczek - Tufts
An Event-driven service environment to enable preservation activities

• Event
  - A significant occurrence that implies the performance of a task (e.g., preservation task)
  - An action that involves at least one object, agent, and/or rights entity (PREMIS)

• Event Outcome
  - A situation or state that follows an event and is a result of the event
Preservation Services Architecture

Preservation Portal

Other Preservation Services
Alerting  Migration Reporting  Statistics . . .

Preservation Monitoring  Messaging  Preservation Integrity

Fedora Repository Service
Digital Objects  Object Content Models

Fedora Service Framework
Format Registry
Event Messages
(services produce and consume them)

Services and Applications
- Message Producers
  - Crawler
  - Fedora Repo
  - Other

Message Consumers
- Fedora Messaging Service
- Preservation Services
  - Preservation Integrity
  - Preservation Alerting

Event Msg Structure
- Header
- Properties
- Body (payload)

PREMIS event
Fedora Event Management

- **Messaging Service in Fedora Framework**
  - Messages represent *events* with actions and outcomes
  - Fedora will provide a *middleware messaging* solution
    - based on open-source Java Messaging Service (JMS)
  - Any *service* can *produce* and/or *consume* event messages

- **Fedora Working Group Focus**
  - The event message format will be based on the PREMIS event
  - Initial event types: ingest, delete, modify, fixityCheck
Event Management – Ingest Example

User Input

Workflow Management System

JMS Message Queue

<eventType>ingest<>
<eventType>modify<>
<eventType>fixityCheck<>
<eventType>
<eventType>

Preservation Integrity Service

Preservation Alerting Service

Fedora Repository Service

Digital Object Ingest

XML

JMS (snd/rcv)

JMS (snd/rcv)

JMS (snd/rcv)
Roadmap for new preservation services

- **Fedora Development Team**
  - Checksums on content bytestreams (Q1 2007)
  - Messaging service in Fedora framework (Q1 2007)
  - Formal expression and registration of “content models” (Q2 2007)
  - Object validation based on “content models” (Q4 2007)

- **Sun Center of Excellence Partnership - Rutgers University Library**
  - Preservation services (2007)

- **Community Development**
  
  Our goal is to enable the Fedora community in developing other preservation features and services
Inputs to Our Work

- RLG/NARA draft “An Audit Checklist for the Certification of Trusted Digital Repositories”


- OAIS Reference Model


Fedora Web Site
www.fedora.info

Fedora Wiki
http://www.fedora.info/wiki