Fedora Repository Tutorial
OA16
Geneva, 17 June 2009

Edwin Shin
Sr. Developer, Fedora Commons
Technical Director, MediaShelf
Target Use Cases

Scholarly Communication

Data Curation and Linking

Science

Humanities

Semantic Knowledge Spaces

Preservation and Archiving

R.D.W. Connor
Current* Users (127)

- University Libraries and Archives - 55
- National Libraries and Archives – 16
- Corporations – 12
- Research Groups and Projects – 11
- Virtual Digital Libraries - 6
- University Information Technology – 6
- Repository Consortia – 5
- Publishing – 4
- Museums and Cultural Organizations – 4
- Medical Centers and Libraries – 4
- Government Agencies – 2
- Professional Societies – 2

* current as of early 2009
Highlighted Projects

- National Science Digital Library
- Bibliothèque nationale de France (BnF): SPAR
- Portuguese National Archives: RODA
- Jewish Women's Archive
- Encyclopedia of Life
- Stanford, Uva, Hull: Hydra Project
- Max Planck Institute
- Public Library of Science
Overview & Background

- Digital Object Model
- Core Repository Service
- Service Framework
Digital Object Model

- Flexible object model supports
  - Documents, articles, journals
  - Electronic Scholarly Texts
  - Digital Images
  - Complex multimedia publications
  - Datasets
  - Metadata
  - Learning objects
  - More…
Fedora Digital Object

- Aggregate content "datastreams" … any type of content
- Intermix both local and external content
- Relationships among digital objects (via RDF)
- Register "content models" for known object patterns

Persistent ID

- DC
- RELS-EXT
- Audit Trail
- Policy

Reserved Datastreams

Datastreams
(any type, any number)
Forming Collections with Relationships

```
PID1

Query

isMemberOfCollection

PID 2

isMemberOfCollection

PID 3

isMemberOfCollection

PID 4

“Smiley Stuff “
Collection Object
```
Digital Objects: Atomistic Model

- Persistent ID
  - Descriptive Metadata
  - Preservation Metadata
  - Image

- Persistent ID
  - Descriptive Metadata
  - Preservation Metadata
  - Image

- Full Text
Content Models

- Establishes a uniform way to classify objects
- Provides a uniform way to access the model
- Enables sharing content and service designs, validating objects
- Enables adding customized functionality to content and sharing services
Services for Digital Objects

What: “Micro services” for digital objects
- Expose extra sets of web-accessible end points on digital objects
- Provide different views or transformations on objects
- Can be easily deployed upon digital objects that have content models associated with them

Why:
- Interoperability: associate a common interface (set of web-accessible end points) to normalize access to heterogeneous digital objects
- Extensibility: at any time, associate a new interface to expose new ways to access content
Core Repository Service

- Modular
  - RDF-based indexing (semantic triplestore index with query)
  - Security with pluggable authentication and XACML policies
  - Journaling (redundancy & failover)
  - Messaging (JMS)
  - Akubra (persistent storage layer)
- Web service interfaces (REST/SOAP)
- Versioning
- Dynamic service binding based on object content model types
- File-centric (all essential characteristics in XML files)
• Simple replication of repositories
  • Replica repositories, each with different underlying storage
  • Useful for failover, redundancy, archiving
Publish and subscribe

- Core repository service can publish API-M events
- Services can subscribe and listen to events
- Services can publish their own events
- Provides a “glue” for the service framework
  - Services can listen for repository events to update themselves
  - Services can listen for events to do a job (e.g., format migration)
Expose the repository as a graph
- Relationships between objects
- Relationships to external entities
- Attributes of objects
- Query the graph

RDF provides a generalizable, extensible data model
- Avoid fixed schema problems and metadata mud wrestling
- Freedom to add and interleave statements from multiple ontologies
- Organic evolution

Powerful queries and inference for repository management
- Transitive relationships among objects
- Dependency analysis
- Detection/Extraction of sub-graphs
- Provenance of disseminations
Security

- Pluggable authentication
  - XML
  - LDAP
  - Shibboleth
- XACML authorization
  - Extremely flexible access control policies
XML Serialization of Fedora Digital Objects

- Ingest and Export formats
  - FOXML
  - METS
  - ATOM
  - ORE (planned)

- Extensible to accommodate new XML formats
  - FOXML is the internal storage format
  - Simple XML format directly expresses Fedora object model
  - Defined by an XML schema
  - Easily translated to other well-known formats
Service Framework

Service APIs

- SOAP: API-A, API-M, “Lite” APIs
- REST-API
  - Ruby & Python bindings
- RISeach: SPO, TQL, SPARQL
- JMS & STOMP

Services

- SWORD, unAPI
- GSearch
- OAI-PMH Provider
- Generic Search Service
  - Lucene
  - Solr
  - Zebra
• A caching, polling OAI-PMH provider
• Supports sets expressed as RDF relationships in RELS-EXTs
Fedora Performance and Scalability Wiki

- 14 million objects (see TestData) with roughly 750 million triples
On the horizon

- Storage: multiplexing stores, DuraCloud
- Resource Index: more index configurability
- Interfaces: AtomPub, WebDAV
- Security: FeSL
- Plugin Architecture: OSGi/Spring DM
Digging deeper

- Kicking the tires
  - Java installer
  - Fedorazon

- Support
  - fedora-users mailing list
  - wiki, tracker
  - vendors: VTLS, MediaShelf

- Training
  - Fedora Users Group meetings
  - RIRI: Red Island Repository Institute
    20-24 July, Prince Edward Island
Questions or comments:
edwin.shin@yourmediashelf.com