DuraCloud DfR

PASIG Austin
January, 2012

Jonathan Markow
Chief Strategy Officer

DuraCloud DfR

A DuraSpace Project Funded by the Alfred P. Sloan Foundation

Why DuraCloud DfR?

- Protect vulnerable research data
- Enable archiving, access, and preservation
- Facilitate cooperation between researcher and institutional data managers
- Provide services to support the research process

Project Structure

- Advisory Group
- Institutional participants
- Technology team

- Facilitated workshops
- Interviews/focus groups
- UX design and development
- Infrastructure development
- Iterative development
- Prototype testing

Advisory Group

Expertise:

- Researcher support
- Support for institutional data management planning
- Repository software and VRE design, development
- Legal, regulatory compliance

Workshop I Participants and Contributors

- Cornell
- George Washington
- Georgia Tech
- Harvard
- ICPSR
- Johns Hopkins
- MIT
- NCAR/UCAR
- Oregon State
- Rice
- Smithsonian
- U of Oregeon
- U of Prince Edward Island
- U of Virginia
- Fluid Project
- DuraSpace

Workshop Outcomes

- Discuss key institutional successes, challenges and priorities in research data management
- Brainstorm requirements and set priorities for DfR
- Discuss project success factors and next steps

Top Five Priorities

#1: Connect the operational and archival phases of the data management lifecycle.

Top Five Priorities

#2: Create simple workflows across the data management lifecycle that automatically capture metadata and provenance.

(…and create incentives for additional metadata creation)

Top Five Priorities

#3: Ensure confidentiality, security, privacy, and predictability of data in the cloud. (Trust and Control)

Top Five Priorities

#4: Automate basic metadata creation and “catalogue” creation.

Top Five Priorities

#5: Create interoperability of operational systems, archiving solutions, and discovery systems used by specific research communities.

Some Other Priorities

- Management of data created beyond the institution
- Controlled sharing of files
- Manage and archive small data sets
- Full life cycle view of storage, backup, replication and archiving.

Some DfR Principles

- Open source, enterprise software solution
- Capture data close to the source
- Don't interfere with researchers' processes
- Provide incentives, added value for metadata creation
- Easy to use; workflows for collaboration, hand-off to institution

Post-Workshop

- List of “user stories” (high level functional spec)
- Project scope (subset)
- Architectural design
- Validate approach with advisors, institutional participants
- Work on UX design
- Forge partnerships with collaborators

Architecture

Take advantage of open source reuse wherever possible, e.g.:

- Authentication (Enterprise SSO, SAML2)
- Cloud-based repository components
- Encryption, service bus, messaging
- DMP tools, persistent identifiers, citations, publishing references,
- etc.

Components

Authentication:

- Pluggable
- Shibboleth support in 1.0 release

Components

DuraCloud

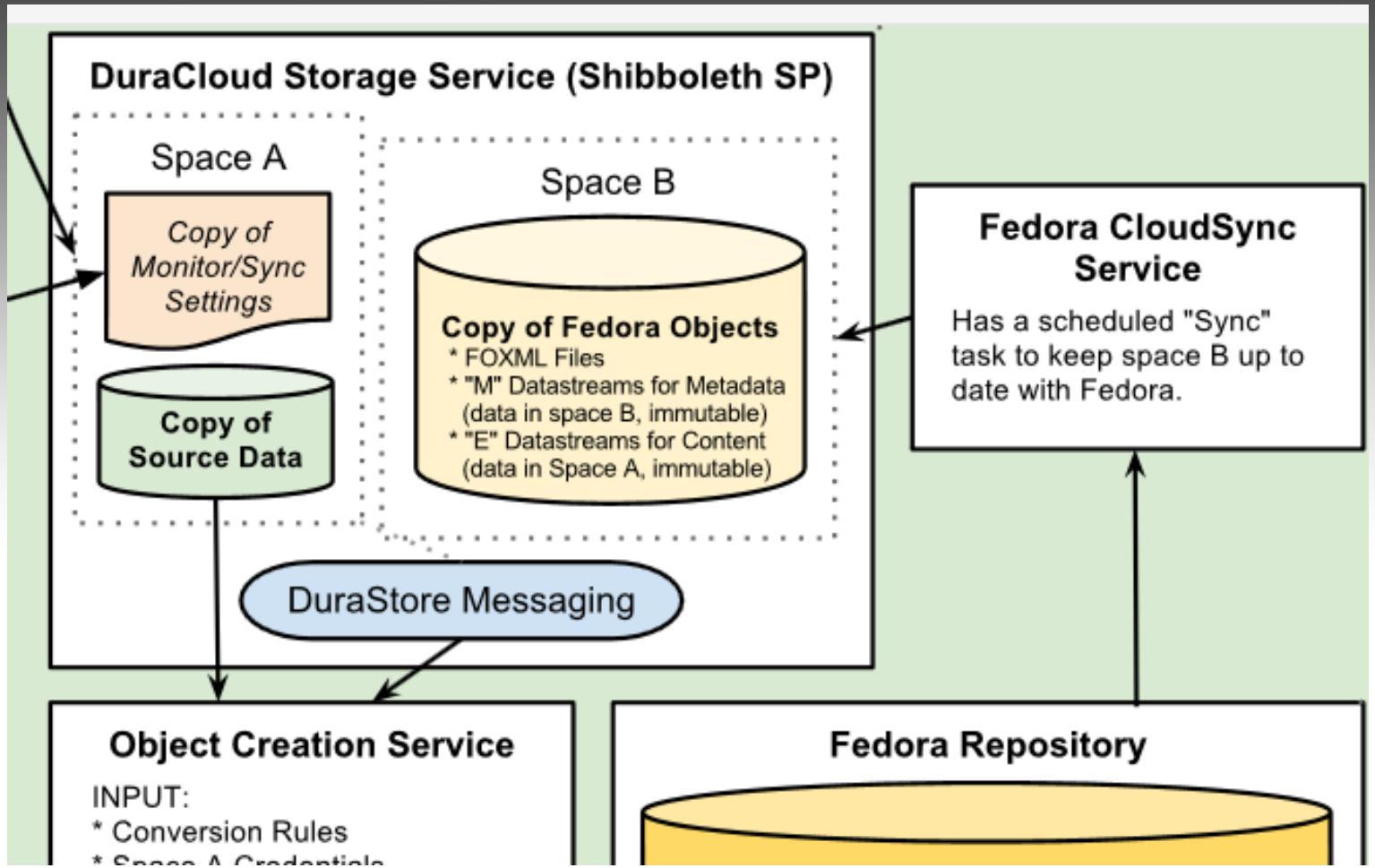
- Synchronization with existing researcher operational systems via “Monitor/Sync” service

Components

Cloud-based Fedora:

- “Object Creation Service” responds to DuraCloud messages, creates Fedora objects
- CloudSync saves Fedora objects back to DuraCloud

Components


DuraCloud Storage Service (Shibboleth SP)
Space A
Copy of Monitor/Sync Settings
Copy of Source Data
Space B
Copy of Fedora Objects
* FOXML Files
* "M" Datastreams for Metadata (data in space B, immutable)
* "E" Datastreams for Content (data in Space A, immutable)
DuraStore Messaging
Fedora CloudSync Service
Has a scheduled "Sync" task to keep space B up to date with Fedora.
Object Creation Service
INPUT:
* Conversion Rules
Fedora Repository

User eXperience

- Reuse existing tools for visualization, manipulation of research data
- Create seamless experience for researcher
- Utilize best practices for interaction design

Partnerships

- Smithsonian Institution – UI for the management and visualization of research project data
- Internet2/InCommon (in discussion)
- Fluid Project – User Interaction Design (www.fluidproject.org)

Project Schedule

- Iterative development
- Evolving prototypes
- User Focus
- Expanding group of participants
- First production release at end of 2012 (Interim releases throughout the year)

Questions?

DuraSpace: www.duraspace.org

DuraCloud: www.duracloud.org

DuraCloud DfR Project:
https://wiki.duraspace.org/x/ZBfNAQ

jjmarkow@duraspace.org