Moving BIG DATA Into DSpace using SWORD & GLOBUS

Lee Taylor, University of Exeter, UK
11 July 2013
Exeter Landscape 2011

- DSpace in use for three independent repositories
- A brand new ~£1M Petabyte data store established to hold all completed research data
- Huge demand but no obvious solution to ingest large data sets some of which of the order of TBs
- Some DSpace developer experience across IT but patchy and part time
OpenExeter Project

● 18 Month JISC funded project looking at Human Factors in Research Data Management
● Early findings suggested research data widely distributed & often on personal PCs off campus etc
● Technical strand focused on establishing Research Data Archive and pulling in the data to a new combined DSpace repository
● DSpace developer resource seconded 80% time
Dsparse not really designed to take BIG DATA?

- Key limitation of DSpace UI is real time upload of data via http – not feasible for TBs of data
- Some support for transfer of data direct to filesystem and link to submission metadata at an administrative level – “submit by reference”
- Look to exploit this feature by programmatically transferring data independently of DSpace and augmenting the built in submission method
The Globus Connection

- GlobusOnline is a free to use cloud based service for data transfer
- Developed by the University of Chicago & Argonne National Labs
- Based on proven GridFTP protocol and used by researchers worldwide for over a decade
- Highly efficient, resilient, secure, open API, “fire & forget”
- Data transfer is peer to peer and does NOT flow through Globus – it merely orchestrates transfer

Globus 1 Minute Overview YouTube
Initial Globus Limitations

- All users must sign up personally to the Globus service but we did not want to manage user identities outside of the normal Exeter authentication and SSO systems.
- Service needed to look & feel like it was part of our Exeter service.
- By default all transfers are “owned” by the logged in user and could only be monitored for completion by this user.
Solutions

- Globus worked with us from the start to understand our requirements and create new functionality where needed
- Authentication – key breakthrough with OA4MP and Exeter SSO system with a big helping hand from Jim Basney & the team at CILogon
- Globus provides us with a branded web site for sign up and general file transfer
- Modified model to allow owner of a destination endpoint to monitor transfers to that endpoint
SWORD Missing Link

- Engaged with one of SWORD authors, Richard Jones of Cottage Labs to update SWORD with capability to support “submit by reference” with DSpace
- With Richard’s help we mapped out what we were trying to achieve in a sequence chart and identified key elements for development
- Crucially we needed a new method within SWORD which allowed us to ingest independent of file transfer
Webapps authenticate using LDAP service accounts which are mapped into Globus and DSpace
Basic Use Case

- End user logs into repository using SSO
- Starts a submission and must register with Globus if this is their first time
- Is automatically logged into Globus and the submission tool (SSO)
- Chooses a “Collection” and enters required metadata for that collection
- Creates a new Globus endpoint if required
- Selects any existing Globus endpoint
- Selects files/directories for transfer
- Logs out and is notified of progress via email
High Level Architecture

Submission Tools

Single Sign On

OAuth
MyProxy

Globus Online

Research Data

Submissions

DSpace

ATMOS IFS

ATMOS
Thanks for your attention

l.w.taylor@exeter.ac.uk
https://ore.exeter.ac.uk

Time for quick demo ?