Plugging The BIG DATA Gap In DSpace Using SWORD And Globus

Lee Taylor

April 2013

£1m investment in 1PB EMC Atmos Data Store

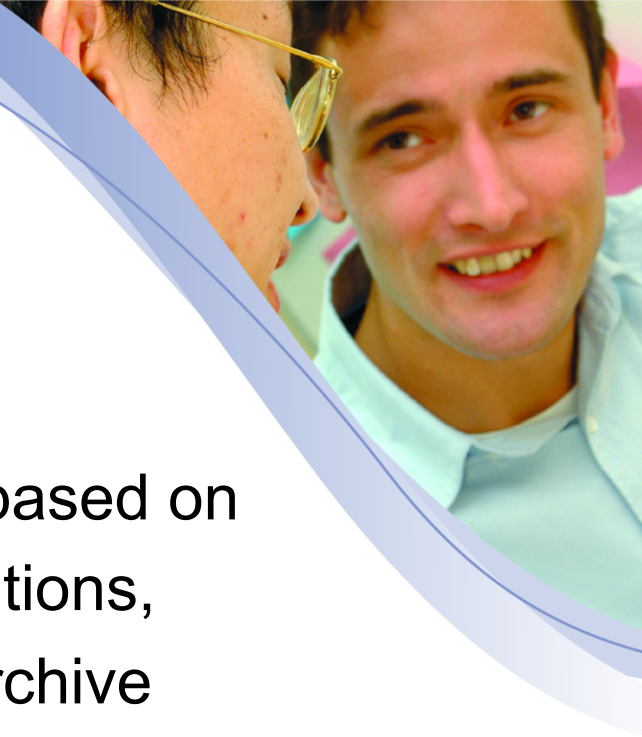
- Atmos is an industry leading data storage solution
- Low cost ~£250 per TB per year
- Object based private cloud storage capable of presenting an Amazon S3 interface + traditional unix like filesystem
- Some limitations due to overhead of metadata per file in object based system

UNIVERSITY OF EXETER

OpenExeter Project

- 18 Month JISC funded project looking at Human Factors in Research Data Management
- Not just science data
- Technical strand focused on use of Atmos for Research Data Archive led by my team
- Early findings suggested research data widely distributed and often on personal PCs off campus
- Key technical aim was to get completed data into DSpace repository for Open Access

UNIVERSITY OF EXETER

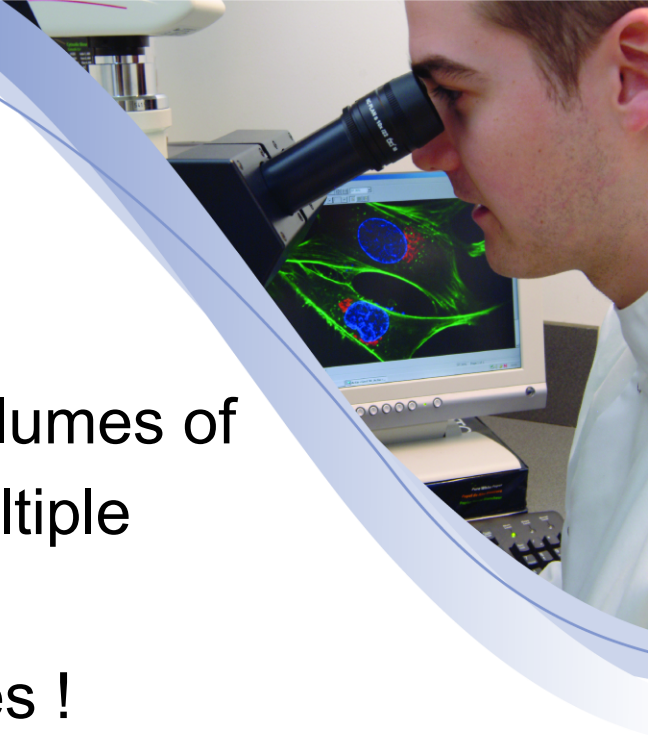


DSpace open source repository of choice

- Aim to combine 3 existing repositories based on DSpace hosting theses, research publications, digital images etc with a research data archive
- 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”

UNIVERSITY OF EXETER

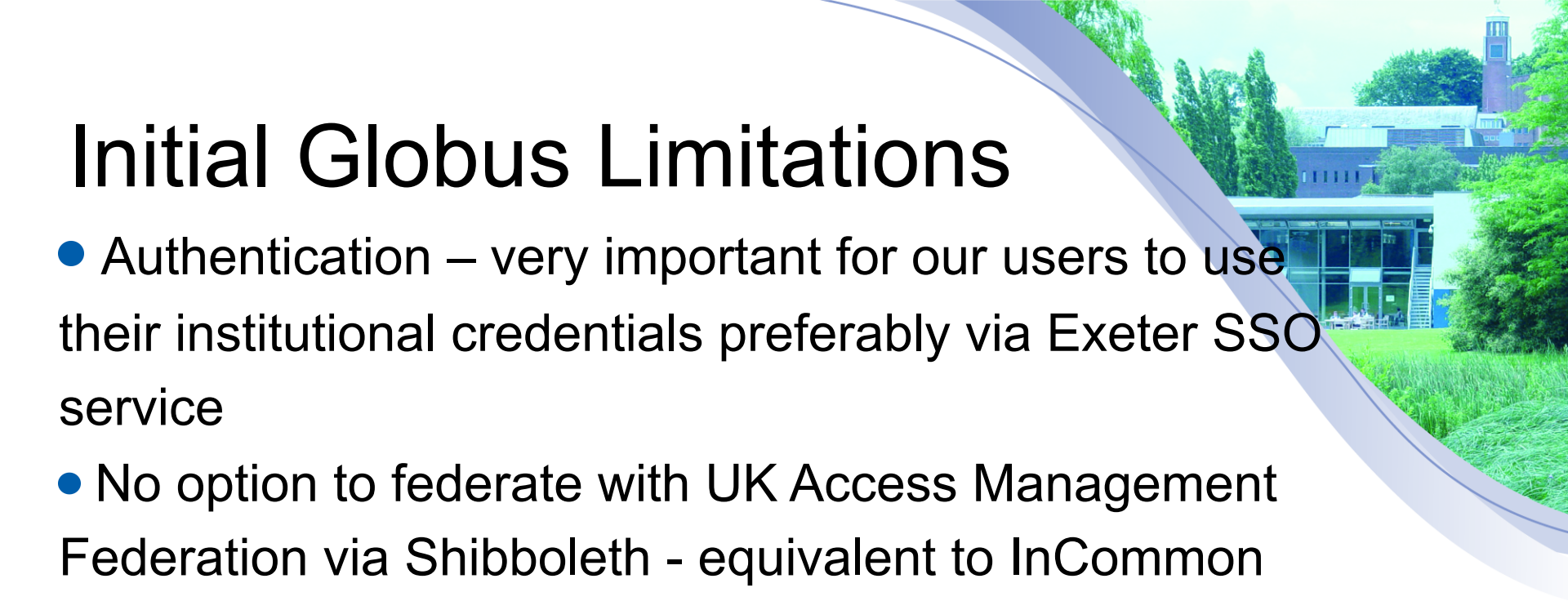
The Globus Connection



- Seeking a way of transferring large volumes of data reliably into the repository from multiple locations
- Globus starts ticking some of the boxes !
- Free (ish), efficient, resilient, secure, open API, cloud service
- Based on proven GridFTP and trusted by researchers worldwide

UNIVERSITY OF EXETER

Initial Globus Limitations



- Authentication – very important for our users to use their institutional credentials preferably via Exeter SSO service
- No option to federate with UK Access Management Federation via Shibboleth - equivalent to InCommon
- Service needs to look & feel like it is part of our DSpace repository
- Ability to monitor transfers for all our users rather than just the transfer owner so that submission is completed in the background

UNIVERSITY OF EXETER

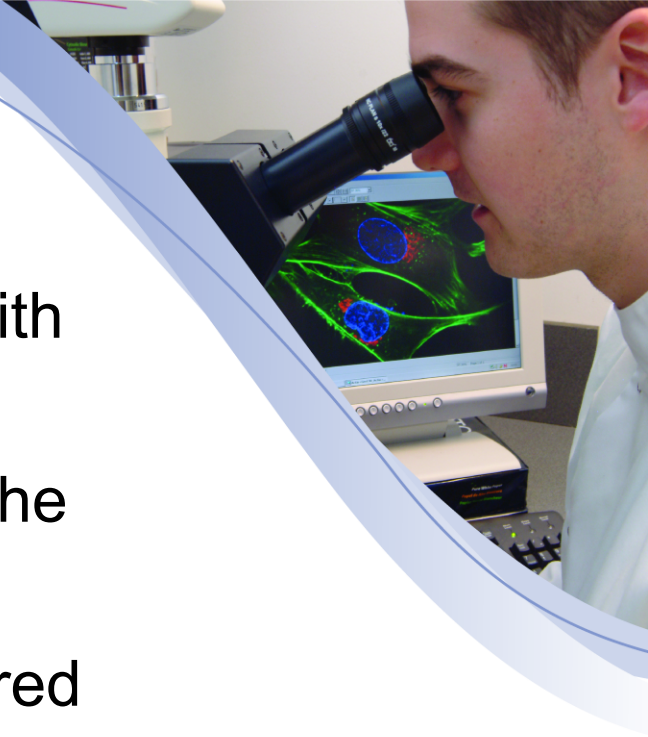
Solutions

- 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
- Some local customisations of DSpace and SWORD now being fed back to the community

Role of SWORD

- SWORD is a lightweight protocol for depositing content from one location to another
- Repository agnostic, open source, largely funded by JISC in the UK
- Engaged with one of SWORD authors, Richard Jones of Cottage Labs to update SWORD with capability to support “submit by reference” with DSpace
- Enables programmatic selection of DSpace collections and item submission

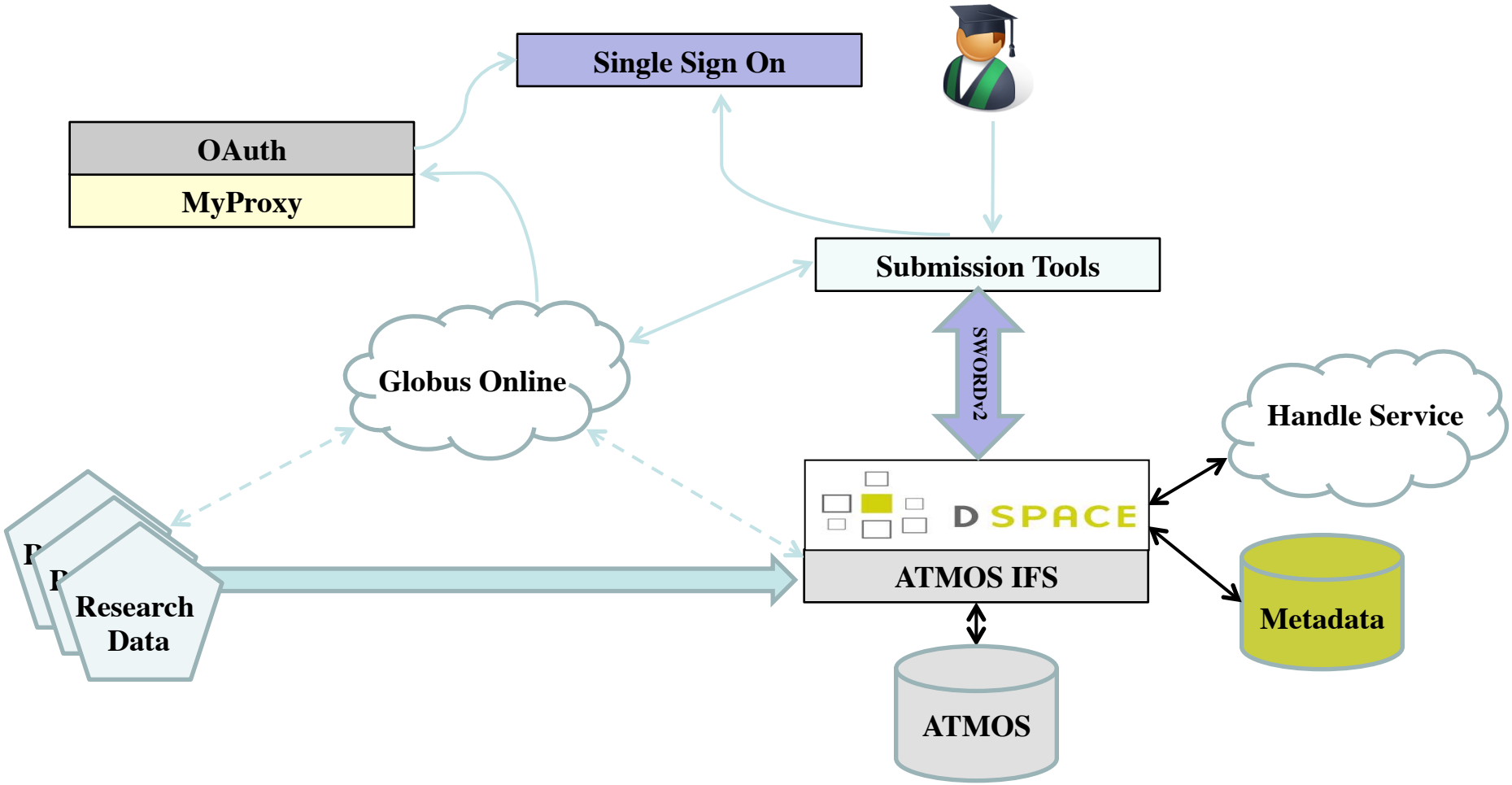
Basic Use Case



- 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 endpoint if required
- Selects an endpoint
- Selects files/directories for transfer
- Logs out and is notified of progress via email

UNIVERSITY OF EXETER

Architecture



Thanks for your attention

Any questions ?