Mission

DRI is a trusted digital repository for Humanities and Social Sciences Data

- linking and preserving the rich data held by Irish institutions, with a central internet access point

- Our Cultural & Social Heritage
Digital Repository of Ireland
Taisclann Dhigiteach na hÉireann

DRI Platform

Access
Preservation
Discovery

Federated Archives, Storage

Linked Logainm

App
App
App
Metadata

Fig. 5: Metadata standards

- Dublin Core
- EAD/ISAD(G)
- MODS
- MARC 21
- Other
- VRA
- INSPIRE
- METS
- No Standard

% of Institutions (excluding 'don’t know')
Formats

![ Formats used by institutions ]

- RTF
- Word
- PDF
- WAV
- MP3
- QuickTime
- MPEG
- TIFF
- JPEG
- JPEG 2000
- RAW
- XML
- TEI
- Other

% of Institutions (excluding 'don’t know')
SYSTEM ARCHITECTURE
OAIS model
Open source components
Custom code to join them together
Why Hydra?

• Previous experience with Fedora Commons and DSPACE
• Wanted to use Solr for search
• Hydra provided a framework for combining Fedora and Solr
• Additional benefits
  – Active user community and support
  – Roadmap that matched our plans
  – Move the data models away from the preservation function
  – Rapid development – we are about 18 months ahead of where we thought we would be
DEVELOPMENT PROCESS
Agile Development Methodology

- Requirements driven
- Daily standups
- 2 week code sprints
- 4-6 week milestones
- Continuous testing/integration/deployment
  - Cucumber, rspec and buildbot
March 2014 – internal prototype
From September 2014 – 6 monthly releases
  • Additional features
  • Additional datasets
Upcoming release contains minimum set of features to provide a TDR
Currently preparing an infrastructure report for publication
DATA INGEST
Data Ingest

DRI supports multiple metadata standards and file formats

Data arranged into collections, with defined owner and editor users

Two data ingest paths

• web interface – add single object to collection
• command line – bulk upload of many objects
Data Ingest

Automated pipeline, using resque, for background tasks on ingested objects

- Virus and malware scan
- Checksuming
- Surrogate generation
- DOI minting
- Linked data – logainm
- Triggered events for certain object types
DATA PRESERVATION
Preservation strategy

Multi-site repository
  Dublin and Maynooth (~25km separation)

Asynchronous replication
  Ability to catch errors on the fly

Segregated storage
  Master copies with surrogates for public access
CEPH features

Using CEPH as the underlying storage system

Provides Posix, S3 and Block access
  Using S3 – potential to move to commercial cloud

Tiered storage and multi-site features

Erasure coding to reduce raw storage needs
USER ACCESS
User Access

Primarily through the blacklight search interface

Other routes

• Curated collections and virtual galleries
• Georeferenced data – mapping
• Temporal data – timelines
• User defined collections
• DOI references in papers
User Access

Anonymous and logged in users

Basic user model – search history, favourite objects, user defined collections

Authentication

• Local users – verified by email
• Shibboleth – link to Edugate
Can provide enhanced access to academic users
Search setup

Solr is one of the leading open source search platforms

Digital objects injected into Fedora Commons
Use the Solrizer gem to create the Solr indices

No deep object inspection at present
Such functionality exists in Solr for text and geo
Other options exist for image
http://projecthydra.org