Building a research data catalogue: knowing what you've got and where it is
DataFinder at Oxford

Sally Rumsey
The Bodleian Libraries
University of Oxford
Damaro project

Data management rollout for Oxford
Proposed architecture of University of Oxford’s modular research data infrastructure

- **Data governance & training**
- **Policies**
- **Training**
- **Sustainability**
- **Data Management Planning Tool [DMPOnline project]**

**Internal environment**
- Data Creation & local (dept) management
  - DataStage
    - Ontologies
  - ViDaaS
- Archival data storage and curation
  - DataBank
    - DOI assigned
    - Ontologies
  - Software Store
- Data discovery and dissemination
  - Document repository
    - Institutional repository: ORA
  - Local (dept) DataFinder
    - DOI assigning
    - DataCite Kernal minimum
    - CERIF compliant
    - Ontologies
  - External Data Stores
  - Regional DataFinder
  - Colwiz

**External environment**
- Oxford local Data Stores
- LabTrove
- External Data Stores
- Ingesting and exposing metadata

**Software Store**

**Institutional repository: ORA**

**Internal environment**

**External environment**

**Ingesting and exposing metadata**
DataFinder
The keystone of Oxford’s Research data infrastructure

- Catalogue/Registry
- Metadata only
- Dissemination
- Discovery
- Citation
- Location
- Irrespective of format
- Compliance with funder requirements
- Explanatory
- Reporting & business intelligence
DataFinder Technical Architecture V1.1

Neil Jefferies,
The Bodleian Libraries
DataFinder user journeys: Admin

- Home/Search
- Admin
  - Restricted via user credentials
  - Admin
    - Add/edit sources
    - Admin permissions
    - Add new user form
  - DataReporter
    - DataReporter home
    - Full record
    - DataReporter results
    - Review contribution
Status of DataFinder records

- **Draft** – user working on record
- **Submitted** – user has submitted record for review
- **Approved** – reviewed submission approved without modification
- **Escalated** – reviewed submission to be checked by other staff due to issues such as commercial or legal agreements, ethics. Note of problem added to admin record.
- **Referred** – more/better information needed before submission can be approved. Submission returned to the submitter with a note of the problem and how to rectify it
- **Rejected** – the admin reviewing the record has decided that there is something fundamentally wrong with it – Reasons for rejection sent to submitter
Welcome to DataFinder

Browse, search, discover metadata from disparate sources from the web using DataFinder and gain access to the data.

DataFinder is the hub of the technical infrastructure that will aggregate metadata from disparate sources, provide dataset registration (through DOIs), and discovery, and location and access details. You can use DataFinder to browse, search, discover metadata at any time, from anywhere on the web. You can accomplish these tasks using a simple web interface of a hosted instance of DataFinder.

Using the web interface

To get started, read this document on the DataFinder wiki
DataFinder Metadata

• RDF schema being written
• Based on DataCite
  – Minimum core
• Export: RDF; CERIF
• Compliance with funders’ requirements
• Subject-specific metadata can be added as .xml stream
• Metadata automatically imported from DataBank
• Manual/edited and harvested kept separate
• Editing: No push back to source
• Harvesting – you get what you’re given
<table>
<thead>
<tr>
<th>Fields in DF web deposit form</th>
<th>Open up on response or +</th>
</tr>
</thead>
<tbody>
<tr>
<td>I should like to deposit data in DataBank</td>
<td>DataBank deposit inc file upload</td>
</tr>
<tr>
<td>Given name of the person(s) who created the data</td>
<td>First name Repeatable</td>
</tr>
<tr>
<td>Family name of the person(s) who created the data</td>
<td>Last name Repeatable</td>
</tr>
<tr>
<td>Indicate the creators' affiliation(s)</td>
<td>Uni; Faculty; Institute; College; Other Repeatable</td>
</tr>
<tr>
<td>Role</td>
<td>Repeatable</td>
</tr>
<tr>
<td>Data owner</td>
<td>Department/research institute</td>
</tr>
<tr>
<td>Is your data an output of funded research?</td>
<td>No No action</td>
</tr>
<tr>
<td>Yes Funding agency (apply mandatory fields per agency) Repeatable</td>
<td></td>
</tr>
<tr>
<td>The title of this dataset to be used in a citation</td>
<td>Sub-title</td>
</tr>
<tr>
<td>Brief description of data</td>
<td>Abstract/description of dataset</td>
</tr>
<tr>
<td>Is your dataset in digital format or not?</td>
<td>Non-digital Location of dataset (email?)</td>
</tr>
<tr>
<td>Digital Location of dataset (UR/DOI/email)</td>
<td>Resource type</td>
</tr>
<tr>
<td></td>
<td>Size</td>
</tr>
<tr>
<td></td>
<td>Format</td>
</tr>
<tr>
<td></td>
<td>Version</td>
</tr>
<tr>
<td>Publication year/Year of release</td>
<td></td>
</tr>
<tr>
<td>Publisher</td>
<td></td>
</tr>
<tr>
<td>Date (other)</td>
<td>Data (coverage) Repeatable</td>
</tr>
<tr>
<td>Date (data collection)</td>
<td>Date (data collection) Repeatable</td>
</tr>
<tr>
<td>What is the earliest date users can access this data?</td>
<td>Open</td>
</tr>
<tr>
<td></td>
<td>Embargoed</td>
</tr>
<tr>
<td></td>
<td>Dark</td>
</tr>
<tr>
<td></td>
<td>Invisible</td>
</tr>
<tr>
<td>Are the data subject to any access T's &amp; C's?</td>
<td>No Access terms and conditions</td>
</tr>
<tr>
<td>Yes Default ODL? + options</td>
<td></td>
</tr>
<tr>
<td>Select subject headings that most closely describe your data</td>
<td>FAST subject headings</td>
</tr>
<tr>
<td>-------------------------------------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Options for other subject class</td>
<td>Repeatable</td>
</tr>
<tr>
<td>Name of subject class scheme</td>
<td>Repeatable</td>
</tr>
<tr>
<td>Subject specific metadata</td>
<td>Repeatable</td>
</tr>
<tr>
<td>Name of subject specific metadata</td>
<td>Repeatable</td>
</tr>
<tr>
<td>Keywords</td>
<td>Repeatable</td>
</tr>
</tbody>
</table>

**Do you have publications associated with this data?**

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Description/abstract of the project</td>
<td>URL/DOI</td>
<td></td>
</tr>
<tr>
<td>Documentation about your data</td>
<td>A short description of the project</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Documentation 1</th>
<th>Text/Link/URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentation 2</td>
<td>Text/Link/URI</td>
</tr>
</tbody>
</table>

| What is the predominant language of the data? | Drop down list |
| Last access request date? |  |
| Rights for metadata | Default. CC0? ODC? |
| Record/digital object ID | UUID |
| Source | Source name |
| Source URL | Source URL |

| Public notes |  |
| Note to reviewer |  |
DataFinder search and browse

- Keyword & Boolean searching
- Filtering by embargo status, subject, source, etc.
- Faceted browsing
- Full record view
- Manual and harvested both retrieved
Locating data

• Discovery points to jump-off page/record
• DOIs will link directly
• Location
  – Digital data: URL, contact details
  – Offline data: contact details
  – Grid data: probably no live link location details
• PURL resolver to ensure links remain live
DataFinder: DataReporter

- Generates standard reports
  - Institutional and departmental reports
  - Click-throughs & downloads
  - Personal data publication reports
  - Records lacking key metadata
  - Statistics for REF
- Customised reports in response to SQL queries
- Admin-only in first instance
Outstanding Challenges

• Culture change
  – Academics
  – PGRs
• Obtaining rich, high quality metadata
  – Aim for good enough
• De-duplication
• Sustainability
• Measuring impact
  – ROI only apparent over many years
• Integrating with other systems eg Oxford DMPOnline
Bigger plan needed

Sally Rumsey
sally.rumsey@bodleian.ox.ac.uk