Research Data Management Infrastructure at Oxford

Neil Jefferies
R&D Project Manager, Bodleian Libraries
Neil.Jefferies@bodleian.ox.ac.uk
A bit of background...
RDM - UK National Context

- Several well-established subject-specific data repositories – e.g. UKDA
  - But, notably, AHDS de-funded
- Growing awareness of importance of RDM throughout first decade of the century
- UK Research councils introducing requirements for data management plans
  - But historically, verification patchy
- 2004 Digital Curation Centre launched
- 2007 UKOLN report for the JISC:
  - ‘Each higher education institution should implement an institutional Data Management, Preservation and Sharing Policy, which recommends data deposit in an appropriate open access data repository and/or data centre where these exist’
- 2009 JISC commences Managing Research Data programme
- 2009 Nature article – ‘Data’s Shameful Neglect’
- 2011 RCUK Common Principles on Data Policy
- 2011 EPSRC Policy framework on Research Data
‘The University of Oxford is committed to supporting researchers in appropriate curation and preservation of their research data, and where applicable in accordance with the research funders’ requirements. It recognises that this must be achieved through the deployment of a federated institutional data repository. This repository has to be supported by a suitable business model, and where possible funded through full economic cost recovery, so that the University can guarantee that the data deposited there will be managed over the long term. The data repository will be a cross-agency activity developed and supported by a number of departments within the University and will build, as far as possible, on existing services, including the Oxford University Research Archive (ORA). It will be overseen by a Steering Group which reports to the University Research Committee. The management and curation of research data will be addressed in cooperation with specialist agencies, research funders and other institutions in the UK and internationally.’
RDM - Oxford History

- **2008** Computing Services internal scoping study into data management requirements
- **2008** Libraries set up DataBank adjunct to ORA
- **2009-10** EIDCSR (Embedding Institutional Data Curation Services in Research)
  - OUCS, OULS, OeRC, Research Services, Computational Biology, Cardiac Mechano-Electric Feedback Group (JISC Funded)
  - Policy, processes, requirements
- **JISC/HEFCE (Universities Modernisation Fund) Projects**
  - **2010-12** Sudamih/ViDaaS – Prototype/productionise Database-as-a-Service
    - Computing Services
    - ORDS (Oxford Research Data Service)
  - **2010-12** Admiral/DataFlow – Prototype/productionise DataStage/DataBank
    - Libraries, Computing Services, OeRC, IBRG, UKOLN, Canonical
    - Lightweight data management/archiving
- **DaMaRO (Data Management Rollout at Oxford)**
  - Integration, Training, Policy (JISC Funded)
  - DataFinder data catalogue
- Draft University Research Data Management Policy
- RDM Portal
- ‘Work Bench’ 3D Image visualisation software
- Initial core RDM metadata schema (being revised)
- Digital curation workflow module, with metadata and archiving client
  - DataFlow progenitor
ORDS – Expunging MSAccess
“Sheer Curation”
- Minimal metadata required
- Enhancement supported

Lightweight, low-impact data management

Network drive & Web UI
- Simple permissions: Personal/group/world

Designed for local or cloud deployment
- Leverage existing infrastructure
- Debian packages/OVF

SWORD2 deposit into DataBank (or anything else!)
DataBank

- Bodleian Data Repository (in dev since 2008) parallels ORA
- “Data” currently defined as “Research outputs that don't fit in ORA”
- File and metadata format agnostic – supports packages (zip & tar) – component subaddressing
- Built on “FEDORA-Lite” object model
- Assigns DataCite DOI's
- Manages embargos – Secure, dark archive is segregated
- Manual and SWORD2 deposit
- REST API
- Debian Packages or OVF

Data Management Roll-out at Oxford (DaMaRO)
Oxford DMPOnline

- Based on DCC DMPOnline tool
- Create, save, submit and use data management plans
  - To accompany research grant applications
  - 20Q's guide the management and publication of data
- Develop a simple DataCite- and CERIF-compliant Data Management Ontology
- DMP's archived in Oxford DMPBank instance of the DataBank software
- Captures metadata in advance of data deposit

JISC

Data Management Roll-out at Oxford (DaMaRO)
The DaMaRo Project
The DaMaRo Project
Diversity is the Key Challenge

- Data management practice differs between disciplines
  - Some don't consider their material to be data
  - Training and education to bridge the gap

- Data is not and will never be located in the same place
  - DataBank, Subject repositories, Grid, offline, non-digital
  - Cataloguing & discovery but also acquisition, accession and forensics may be needed

- Metadata standards development and adoption varies widely
  - Bioinformatics boasts 200+ standards for describing experiments
  - Tools like Elastic Search are essential
  - Support domain specific applications built over archives
  - Standards development and promotion at the other end of the spectrum

- Data retention and metadata requirements vary
  - Funders mandates vs unfunded research
  - Legal requirements (IPR vs FOI)
  - Citation requirements (DataCite)

- Interoperability
  - Research Information Management (CERIF)
  - Research communities (Linked Open Data)
  - Libraries and Archives (OAI-XXX, SWORD2)

Data Management Roll-out at Oxford (DaMaRO)
Data Management Roll-out at Oxford (DaMaRO)
DataFinder

- Catalogue/registry of research data
  - Wherever and whatever it is!
  - OAI-PMH harvesting of external data stores
  - Manual record entry for non-electronic or non-harvestable data
- Search/browse interface
- DataReporter module
  - CERIF compatible
  - Analytics as well as content statistics
- Core Metadata schema based on DataCite
- Interfaces with many systems
  - “Hub” Of RDM activity
- Hierarchical architecture
  - Local catalogues, subjects specific or inter-institutional catalogues possible
It lives!

Data Management Roll-out at Oxford (DaMaRO)
Metadata (again)

- **Citation**
  - DataCite kernel: Creator, Title, Date, Publisher*, ID*

- **Discovery**
  - The more the merrier. Domain specific metadata is great (if not very tractable)

- **Funder requirements**
  - EPSRC: “Sufficient metadata should be recorded and made openly available to enable other researchers to understand the potential for further research and re-use of the data”
  - Meh!

- **Assessment of usefulness/value**

- **Preservation**
  - Some can be autogenerated
  - File format diversity can be a challenge

- **Reporting and Business Intelligence**
  - Different standards like CERIF require crosswalks/mappings

- **Manual entry generally disliked**
  - Import from existing systems (other repositories/research platforms)
  - Acquire from researcher interactions with other systems (DMP, Datastage, ORDS)
## Minimum Core Data (WIP!)

<table>
<thead>
<tr>
<th>Element</th>
<th>Auto Gen</th>
<th>DataCite</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record/digital object ID</td>
<td>UUID</td>
<td>M</td>
<td>If no URL: contact details</td>
</tr>
<tr>
<td>Location of dataset</td>
<td>URL/DOI</td>
<td>DataBank auto</td>
<td>To enable indication of non-digital data; Check box + options. On/offline</td>
</tr>
<tr>
<td>[Medium]</td>
<td>Default: digital (+ non-digital)</td>
<td></td>
<td>To enable indication of non-digital data; Check box + options. On/offline</td>
</tr>
<tr>
<td>Creator (if not depositor)</td>
<td>Repeatable</td>
<td>WebAuth/OxDMP</td>
<td>If depositor draw from WebAuth; (see optional)</td>
</tr>
<tr>
<td>Creator affiliation (if not depositor)</td>
<td>Repeatable (see optional)</td>
<td>WebAuth/OxDMP</td>
<td>If depositor draw from WebAuth; CUD; Imply subject</td>
</tr>
<tr>
<td>Title</td>
<td></td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>Publisher of data</td>
<td>Default University of Oxford</td>
<td>Default</td>
<td>If an embargo period has been in effect, use the date when the embargo period ends.</td>
</tr>
<tr>
<td>Publication year</td>
<td>Default current</td>
<td>Default</td>
<td></td>
</tr>
<tr>
<td>Access terms &amp; conditions</td>
<td>Default + options</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data owner</td>
<td>Default Department</td>
<td>WebAuth/OxDMP</td>
<td>For curation; ALT Name (Person or role) + Data owner contact. + Qu 'Do you own the rights for this data?' Need policy</td>
</tr>
<tr>
<td>Access date to data</td>
<td>Default current</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rights for metadata</td>
<td>Default: CCO? ODC?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Subject]</td>
<td>FAST + options</td>
<td></td>
<td>Import where possible using available data. Encourage imupt.+ K/w option. See Optional</td>
</tr>
</tbody>
</table>

---

Data Management Roll-out at Oxford (DaMaRO)
### Context Dependent Mandatory Metadata (WIP!)

<table>
<thead>
<tr>
<th>Element</th>
<th>Auto Gen</th>
<th>DataCite</th>
<th>EPSRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Funding agency</td>
<td>OxDMP</td>
<td></td>
<td>M</td>
</tr>
<tr>
<td>Grant number</td>
<td>OxDMP</td>
<td></td>
<td>M</td>
</tr>
<tr>
<td>Project information</td>
<td>Link to project web page/blog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last access request date</td>
<td>Automatically determined</td>
<td></td>
<td>M</td>
</tr>
<tr>
<td>Source</td>
<td>Imported record</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source URL</td>
<td>Imported record</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data generation process</td>
<td>Text or link to paper/document</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Why the data was generated/Abstract/Brief description</td>
<td>Might be link to project page</td>
<td></td>
<td>M</td>
</tr>
<tr>
<td>Date</td>
<td>Repeatable; eg date (range) of data collection; format described in W3CDTF</td>
<td>0</td>
<td>M</td>
</tr>
<tr>
<td>Reason for embargo</td>
<td>Repeatable; List options</td>
<td></td>
<td>[M]</td>
</tr>
</tbody>
</table>
Where Next?

- **Oxford DAMASC (Databank Archiving and Manuscript Submission Combined)**
  - Bodleian and OUP: Data deposit into institutional data archive alongside publisher paper submission workflow with cross citation

- **Author identification project**
  - Identity management across Libraries, CRIS, Publishers etc.
  - Based on sameas service – there will never be a single standard!
  - Privacy concerns

- **ViDaaS, DataBank and DataStage generating interest at a number of institutions**
  - Transition to a more managed Open Source project arrangement
  - Sustainability model needs to be defined
  - Interoperability with wider spectrum of systems

- **DataBank/DataFinder Roadmap**
  - Large file handling – just pass download details at the point of submission
    - File can be acquired asynchronously in the background
  - Group management for DataFinder/DataBank - delegation and group administration
    - Balance simplicity with requirements – challenge of mapping Oxford's org structure

- **Methodological publications (e.g. MyExperiment)**
  - Bridge data and papers
  - Cover case where recreation cheaper than storage
Questions?

Neil Jefferies
R&D Project Manager, Bodleain Libraries
Neil.Jefferies@bodleian.ox.ac.uk