Building a Research Data Registry for the UK

Alex Ball\textsuperscript{1}  Kevin Ashley\textsuperscript{2}  Patrick McCann\textsuperscript{3}  Laura Molloy\textsuperscript{3}  Veerle Van den Eynden\textsuperscript{4}

\textsuperscript{1}DCC/UKOLN Informatics, University of Bath  
\textsuperscript{2}DCC, University of Edinburgh  
\textsuperscript{3}DCC/HATII, University of Glasgow  
\textsuperscript{4}UK Data Archive

12 June 2014
Outline

Motivation

Previous work

Architecture

Collaborators

Metadata

Evaluation

Phase 2

Conclusions
UK data landscape
Motivation

- **Researchers** need to search efficiently across many repositories.
  - Not just specialist data centres any more...  
    * Institutional data repositories  
    * Generalist repositories related to journals  
  - Interdisciplinary and multidisciplinary research requires data drawn from diverse sources.

- Funders need to track data assets funded by their grants.
- University administrators need to track data assets produced by their researchers.
- Datasets need more visibility and ‘identity’.
Motivation

- Researchers need to search efficiently across many repositories.
- **Funders** need to track data assets funded by their grants.
  - Impact of data as a first class research output
- **University administrators** need to track data assets produced by their researchers.
  - Impact of data as a first class research output
- Datasets need more visibility and ‘identity’.
Motivation

- Researchers need to search efficiently across many repositories.
- Funders need to track data assets funded by their grants.
- University administrators need to track data assets produced by their researchers.
- **Datasets** need more visibility and ‘identity’.
  - Data sharing is good for research, but held back by
    - Low levels of reuse
    - Variety of attribution mechanisms
    - Lack of formal impact measurements
Researchers need to search efficiently across many repositories.

Funders need to track data assets funded by their grants.

University administrators need to track data assets produced by their researchers.

Datasets need more visibility and ‘identity’.

We need a Research Data Registry and Discovery Service!
Welcome to the INSPIRE geoportal

The INSPIRE Directive requires the Commission to establish a community geo-portal and the Member States shall provide access to their infrastructures through the geo-portal as well as through any access points they themselves decide to operate.

More...

Discovery / Viewer
Search, discover and access geographic information provided by European governmental, commercial, and non-commercial organizations.

More ...

Validator
The purpose of the INSPIRE Metadata Validator is to test the compliancy of INSPIRE metadata with the INSPIRE Metadata Regulation.

More ...

Metadata Editor
Create metadata according to the INSPIRE implementing rules.

More ...
Search for data

Please use one or more of the options below to search the catalogue. Alternatively, try our advanced text search.

Free text

Please enter your search terms below or leave blank to match everything (help)

Search

Date range

Match data overlapping the period (help)

From

To

Search

Clear

Geographic search

The map display lets you draw a box to define an area of interest and only returns records that intersect the box. We also offer predefined areas below (help)

Countries

Select a country...

Search

Clear area

2130 results returned in 0.16 seconds.

You are searching for...

everything in the catalogue.

Shortcut search

Click to search records for

Atmosphere
Oceans
Biota
Geo-scientific Information
Inland waters
Environment

Only return records with an online resource
The CESSDA Catalogue

- Provides a seamless interface to datasets from social science data archives across Europe
- May be searched via a free text search from the top search box
- Browsed via the options in the left hand side menu
- Can be viewed in any of nine languages. The default language is dependant on the regional setting of the computer user
- The language can be switched at any time by selecting the relevant language from the drop down list above. **Please note:** changing the language will take the user back to this CESSDA Catalogue home page
- **In the free text search:**
  1. * truncation is supported
  2. '?' wildcards are supported
## At a glance

<table>
<thead>
<tr>
<th>Dataset Type</th>
<th>Dataset Language</th>
<th>Funder</th>
<th>Funding Stream</th>
<th>Scientific Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown (611)</td>
<td>English (101)</td>
<td>European Commission (4)</td>
<td>Capacities (4)</td>
<td>INFRA (4)</td>
</tr>
<tr>
<td></td>
<td>German (2)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication Year</th>
<th>Access Mode</th>
<th>Datasource</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013 (1)</td>
<td>UNKNOWN (608)</td>
<td>Datacite (597)</td>
</tr>
<tr>
<td></td>
<td>Open Access (3)</td>
<td>Zenodo (597)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D... (14)</td>
</tr>
</tbody>
</table>
Die hundertjährige Translationsfeier der beiden Wettinger Katakombenheiligen Marianus und Getulius

Integration eines Freiandzeichen-Tools in das Trainings- und Prüfungssystem CaseTrain

Konzeption und Evaluation eines fallbasierten Trainingssystems im universitätsweiten Einsatz (CaseTrain)

Akzeptanz medizinischer Trainingsfälle als Ergänzung zu Vorlesungen

Aufwandsanalyse für computerunterstützte Multiple-Choice Papierklausuren

Die bildlichen Darstellungen im Atlas Marianus des Wilhelm Gumppenberg und eine Wallfahrtsschau in der Bischöflchen Sammlung Freiburg

Data from: Contrasting insights provided by single and multispecies data in a regional comparative phylogeographic study
Spotlight on research data

N.C.W. Beadle Herbarium

The N.C.W. Beadle Herbarium (NE) at University of New England contains around 90,000 pressed, dried, incorporated and databased plant specimens. The collection includes more than 150 TYPE specimens that anchor scientific names as cited in the original publication of those names. This rich resource contains many collections that are of great interest to local and international researchers. The specimen sheet collection of the N.C.W. Beadle Herbarium is databased and available to registered users for online data entry and data query.

Explore the N.C.W. Beadle Herbarium Collection through Research Data Australia >>>
Attractions of the Research Data Australia software:

- Familiar to project team
- Proven technology
- Plays nicely with search engines
- Displays sample citations and access/rights information up front

Challenges of using the software in the UK:

- Not used before outside Australia
- Uses uncommon metadata standard (RIF-CS) internally
- Original implementation only harvests in RIF-CS
- No UK data centre can output RIF-CS metadata
Phase 1 overview

1. Implement a working instance of ORCA.
2. Assemble a group of contributors and establish how their metadata will be harvested.
3. Write crosswalks for transforming contributed metadata into RIF-CS.
4. Harvest metadata from contributors.
5. Reports on
   - using the Research Data Australia software;
   - how harvesting from data centres went;
   - how harvesting from university repositories went;
   - the value of continuing to develop the registry.

(Made available from http://www.dcc.ac.uk/projects/research-data-registry-pilot)
Architecture

http://rdrds.cloudapp.net/

- Access management
- Front end
- Metadata registry
- OAI-PMH harvester
- Indexer (Apache Solr)
- CMS editor
- ID manager

CentOS Linux
MS Azure

Collections without OAI-PMH support

Collections with OAI-PMH support

HTTP

OAI-PMH
Collaborators

Data centres:

▶ UK Data Archive
▶ NERC Data Catalogue Service
  ▶ BADC
  ▶ BODC
  ▶ EIDC
  ▶ NEODC
  ▶ NGDC
  ▶ PDC
  ▶ UKSSDC
▶ Archaeology Data Service

Universities:

▶ Edinburgh
▶ Glasgow
▶ Hull
▶ Lincoln
▶ Leeds
▶ Oxford
▶ Oxford Brookes
▶ St Andrews
▶ Southampton
## Metadata crosswalks

<table>
<thead>
<tr>
<th>DataCite 3</th>
<th>MODS 3.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ Archaeology Data Service</td>
<td>▶ Edinburgh</td>
</tr>
<tr>
<td>▶ Oxford</td>
<td>▶ St Andrews</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DDI Codebook 2.5</th>
<th>OAI-PMH Dublin Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ UK Data Archive</td>
<td>▶ Oxford Brookes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EPrints 3</th>
<th>UK Gemini 2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ Glasgow</td>
<td>▶ NERC Data Catalogue</td>
</tr>
<tr>
<td>▶ Leeds</td>
<td>Service (incl. ADS)</td>
</tr>
<tr>
<td>▶ Southampton</td>
<td></td>
</tr>
</tbody>
</table>
Attitudes of Students at the London School of Economics, February 1980

To conduct a course exercise that collects questionnaire-based information each year from a sample of students at the London School of Economics. The studies focus on background characteristics relevant to a student population, on attitudes to selected political and social issues, and on participation in various activities at LSE. Questions vary somewhat from year to year.

How to Cite this Collection

Citation ( Metadata):
Husbands, C. ( 1 Ja,1 Ja,1 Ja,1 Ja ): Attitudes of Students at the London School of Economics, February 1980.

Access

http://dx.doi.org/10.5255/UKDA-

Access rights

The depositor has specified that registration is required and standard conditions of use apply. The depositor may be informed about usage. See terms and conditions of access for further information.

Connections

People

C. Husbands

Suggested Links

Internal Records
258 records with matching subjects

External Records
1 records from DataCite

Subjects

Keynotes

ABORTION (INDUCED) ALCOHOL CONSUMPTION ATTITUDES
EDUCATIONAL FEES EDUCATIONAL FINANCE EDUCATIONAL GRANTS
FAMILY INFLUENCE FOREIGN STUDENTS GENDER NARCOTIC DRUGS
OCCUPATIONS PARENTS PART-TIME COURSES
POLITICAL PARTICIPATION PORNOGRAPHY SEXUAL BEHAVIOUR
SMOKING SOCIAL ACTIVITIES (LEISURE) SOCIAL CLASS
SOCIAL PROTEST STUDENT HOUSING STUDENT LEISURE
STUDENT PARTICIPATION STUDENTS UNIVERSITY COURSES

Higher and further
Evaluation questions

▶ Does the software work as intended?

▶ Do the harvested records look useful and accurate?

▶ Is the system straightforward to use?

▶ What might be improved?

▶ What additional functions would be desirable?
Evaluation questions

▶ Does the software work as intended?
  Can import metadata, manage records, perform searches, etc.
  OAI-PMH harvest needed work.
▶ Do the harvested records look useful and accurate?

▶ Is the system straightforward to use?

▶ What might be improved?

▶ What additional functions would be desirable?
Evaluation questions

- Does the software work as intended?
  Can import metadata, manage records, perform searches, etc.
  OAI-PMH harvest needed work.

- Do the harvested records look useful and accurate?
  Most records imported at Quality Level 2.
  No systematic problems so far...
  Contributors motivated to optimise OAI-PMH feeds/DataCite submissions

- Is the system straightforward to use?

- What might be improved?

- What additional functions would be desirable?
Evaluation questions

- Does the software work as intended?
  Can import metadata, manage records, perform searches, etc.
  OAI-PMH harvest needed work.

- Do the harvested records look useful and accurate?
  Most records imported at Quality Level 2.
  No systematic problems so far...
  Contributors motivated to optimise OAI-PMH feeds/DataCite submissions

- Is the system straightforward to use?
  Interface could be more streamlined.

- What might be improved?

- What additional functions would be desirable?
Evaluation questions

- **Does the software work as intended?**
  Can import metadata, manage records, perform searches, etc. OAI-PMH harvest needed work.

- **Do the harvested records look useful and accurate?**
  Most records imported at Quality Level 2.
  No systematic problems so far...
  Contributors motivated to optimise OAI-PMH feeds/DataCite submissions

- **Is the system straightforward to use?**
  Interface could be more streamlined.

- **What might be improved?**
  Better to ingest in original format and normalise later?
  Handling of citation data.

- **What additional functions would be desirable?**
Evaluation questions

- Does the software work as intended?
  Can import metadata, manage records, perform searches, etc.
  OAI-PMH harvest needed work.

- Do the harvested records look useful and accurate?
  Most records imported at Quality Level 2.
  No systematic problems so far...
  Contributors motivated to optimise OAI-PMH feeds/DataCite submissions

- Is the system straightforward to use?
  Interface could be more streamlined.

- What might be improved?
  Better to ingest in original format and normalise later?
  Handling of citation data.

- What additional functions would be desirable?
  Richer set of relations supported?
Phase 2

- Define a set of clear use cases.
- Define a set of workflows for using the service.
- Compare different possible platforms for the service and assess their suitability.
- Establish a working instance of the system, involving all UK data centres and university data repositories.
- Establish a simple workflow for adding more data sources to the service, adapting to changes in existing data sources, and avoiding duplication.
- Test the system for usability.
- Produce recommendations for quality and standardisation of metadata records.
- Evaluate the costs and benefits of the system.
Conclusions

- Institutional data repositories need to work harder to make their holdings visible than domain-specific repositories. Cooperating with registries and search engines can help!
- When dealing with metadata it is important to focus on the application you are trying to provide.
- Working with the community is vital to ensure the service is relevant and useful.
- A strong business case and value proposition is important for sustainability.
Thank you for your attention

DCC Website: http://www.dcc.ac.uk/