



D | C | C

because good research needs good data

Building a Research Data Registry for the UK

Alex Ball¹ Kevin Ashley² Patrick McCann³ Laura Molloy³
Veerle Van den Eynden⁴

¹DCC/UKOLN Informatics, University of Bath

²DCC, University of Edinburgh

³DCC/HATII, University of Glasgow

⁴UK Data Archive

12 June 2014



Except where otherwise stated, this work is licensed under the Creative Commons Attribution 4.0 International licence:
<http://creativecommons.org/licenses/by/4.0/>

Supported by



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



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'.

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 ...](#)

```
Invalid Element
Number of inst
(2.2.5) Unique
(1) * identifi
/2005/gmd/
/2005/gmd/
(2.4) For dat
/www.iso/
```

Metadata Editor

Create metadata according to the INSPIRE implementing rules.

[More ...](#)

```
<xsd:sche
<-Doc
<xsd:e
A ADI
```


[About the DCS](#)[Search for data](#)[Results](#)**You are searching for...**

everything in the catalogue.
2130 results returned in **0.16** seconds.

Shortcut search

Click to search records for

- [Atmosphere](#)
- [Oceans](#)
- [Biota](#)
- [Geo-scientific Information](#)
- [Inland waters](#)
- [Environment](#)

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](#))

Only return records with an online resource

Date range

Match data overlapping the period ([help](#))

From

To

Dates should be in the format YYYY-MM-DD or YYYY

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](#))

Define the search box coordinates - input order is most westerly, southerly, easterly, northerly



English

[Help on Searching Data](#) | [Contact](#)

Search

- CESSDA Catalogue
- Browse by Topic
- Browse by Keyword
- Browse by Data Publisher

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

CESSDA Data Publishers

(9140 studies available)

[APIS](#) (27 studies)[ADPSS-Sociodata](#) (59 studies)[CSDA](#) (1009 studies)[DANS](#) (124 studies)[ADP](#) (1087 studies)[FORS](#) (754 studies)[FSD](#) (2010 studies)[NSD](#) (1034 studies)[ISSDA](#) (97 studies)[LIDA](#) (423 studies)[NSD Metadata](#) (1034 studies)[SND](#) (196 studies)[ESSDA](#) (15 studies)[UKDA](#) (642 studies)[GESIS ZACAT](#) (629 studies)

All	Publications	Data	Projects	People	Organizations	Datasources
-----	--------------	------	----------	--------	---------------	-------------

At a glance

Dataset Type

Unknown (611)

Dataset Language

English (101)
German (2)

Funder

European
Commission (4)

Funding Stream

Capacities (4)

Scientific Area

INFRA (4)

Publication Year

2013 (1)

Access Mode

UNKNOWN (608)
Open Access (3)

Datasource

Datacite (597)
Zenodo for Research
D... (14)



beta

Filter

allocator

datacentre

prefix

resourceType

contributor

creator

publicationYear

publisher

language

No active filters. Use the sidebar to filter search results.

10 documents found in 177ms

Page 1 of 1

[Die hundertjährige Translationsfeier der beiden Wettinger Katakombenheiligen Marianus und Getulius](#) # 1

[doi:10.5169/SEALS-115655](#)

Felder, Peter

title: Die hundertjährige Translationsfeier der beiden Wettinger Katakombenheiligen **Marianus** und Getulius

[Integration eines Freihandzeichen-Tools in das Trainings- und Prüfungssystem CaseTrain](#) # 2

[doi:10.3205/10CBT27](#) Text : ConferencePaper

Ifland, Marianus • Hörnlein, Alexander • Ott, Julian • Puppe, Frank

creator: Ifland, **Marianus**

[Konzeption und Evaluation eines fallbasierten Trainingssystems im universitätsweiten Einsatz \(CaseTrain\)](#) # 3

[doi:10.3205/MIBE000086](#) Text : JournalArticle

Hörnlein, Alexander • Ifland, Marianus • Klügl, Peter • Puppe, Frank

creator: Ifland, **Marianus**

[Akzeptanz medizinischer Trainingsfälle als Ergänzung zu Vorlesungen](#) # 4

[doi:10.3205/ZMA000754](#) Text : JournalArticle

Hörnlein, Alexander • Mandel, Alexander • Ifland, Marianus • Lüneberg, Edeltraud • Deckert, Jürgen • (et. al.)

creator: Ifland, **Marianus**

[Aufwandsanalyse für computerunterstützte Multiple-Choice Papierklausuren](#) # 5

[doi:10.3205/ZMA000767](#) Text : JournalArticle

Mandel, Alexander • Hörnlein, Alexander • Ifland, Marianus • Lüneburg, Edeltraud • Deckert, Jürgen • (et. al.)

creator: Ifland, **Marianus**

[Die bildlichen Darstellungen im Atlas Marianus des Wilhelm Gumpenberg und eine Wallfahrtsbilderreihe in der Bischöflichen Sammlung Freiburg](#) # 6

[doi:10.5169/SEALS-339455](#)

Ronner, Christel

title: Die bildlichen Darstellungen im Atlas **Marianus** des Wilhelm Gumpenberg und eine

[Data from: Contrasting insights provided by single and multispecies data in a regional comparative phylogeographic study](#) # 7

[doi:10.5061/DRYAD.M7RC3](#) Dataset : DataPackage



Browse by Subject Area



Browse by Map Coverage



Advanced Search

What's in Research Data Australia



Collections (92633)

Research datasets or collections of research materials.



Parties (25467)

Researchers or research organisations that create or maintain research datasets or collections.



Activities (40674)

Projects or programs that create research datasets or collections.



Services (184)

Services that support the creation or use of research datasets or collections.

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 >>>](#)

Research Data UK?

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 I overview

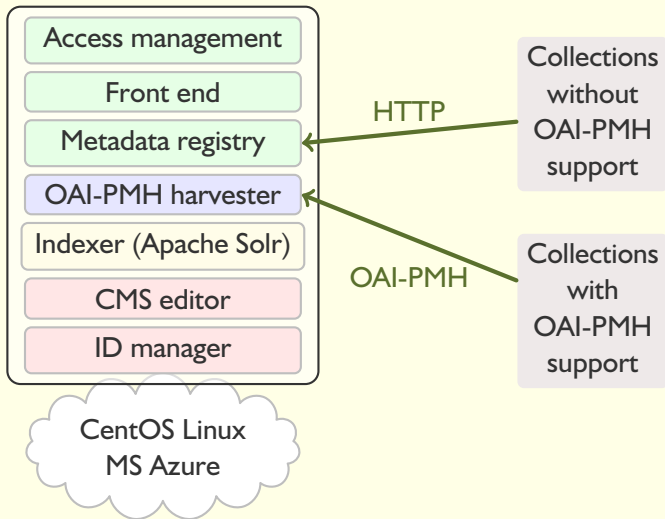
1. Implement a working instance of ORCA.
2. Assemble a group of contributors and establish how their metadata will be harvested.
3. Write crosswalks for transforming contributed metadata into RIF-CS.
4. Harvest metadata from contributors.
5. Reports on
 - ▶ using the Research Data Australia software;
 - ▶ 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/>



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

DataCite 3

- ▶ Archaeology Data Service
- ▶ Oxford

DDI Codebook 2.5

- ▶ UK Data Archive

EPrints 3

- ▶ Glasgow
- ▶ Leeds
- ▶ Southampton

MODS 3.5

- ▶ Edinburgh
- ▶ St Andrews
- ▶ Hull

OAI-PMH Dublin Core

- ▶ Oxford Brookes
- ▶ Lincoln

UK Gemini 2.2

- ▶ NERC Data Catalogue Service (incl. ADS)

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.





because good research needs good data

Thank you for your attention

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

Jisc RDRDS Project: <http://www.dcc.ac.uk/projects/research-data-registry-pilot>

