Metadata for Impact

Lessons from the UK

Alex Ball

DCC/UKOLN, University of Bath

25 September 2014

Supported by Jisc

Outline

Vision

Progress so far

Metadata elements

Conclusions

Acknowledgements

Project team Kevin Ashley, Alex Ball, Patrick McCann, Laura Molloy, Veerle Van den Eynden

Funder Jisc

Scope and purpose

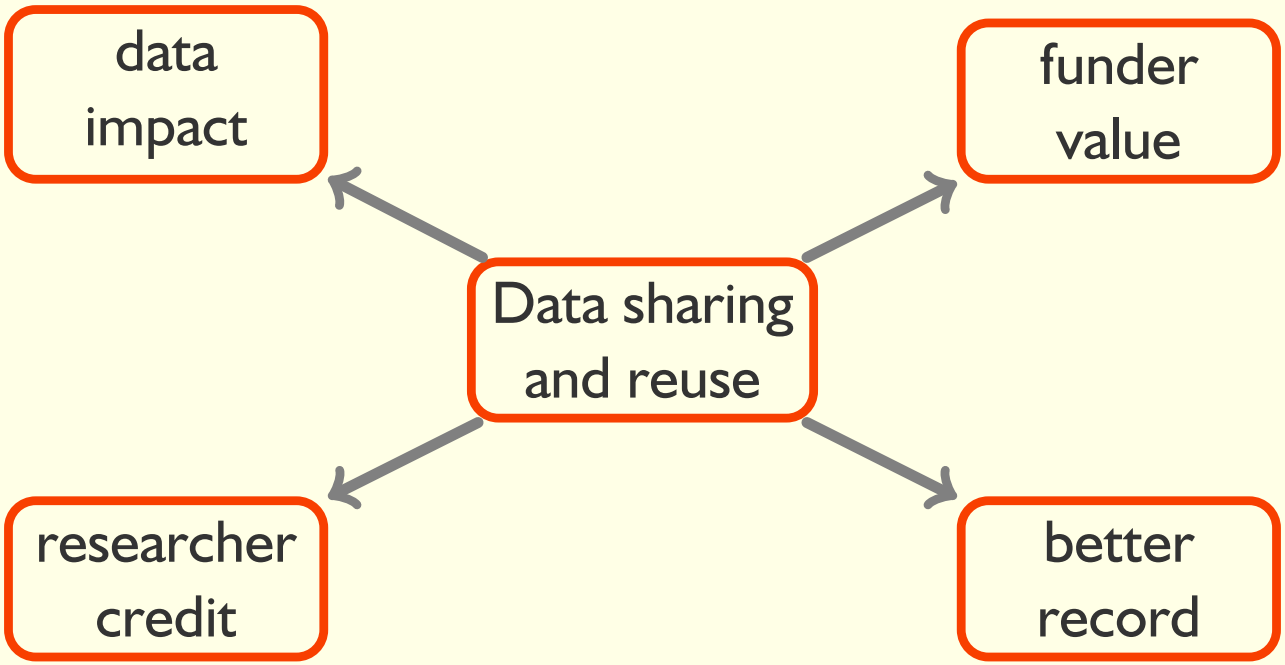
UK Research Data

Scope and purpose

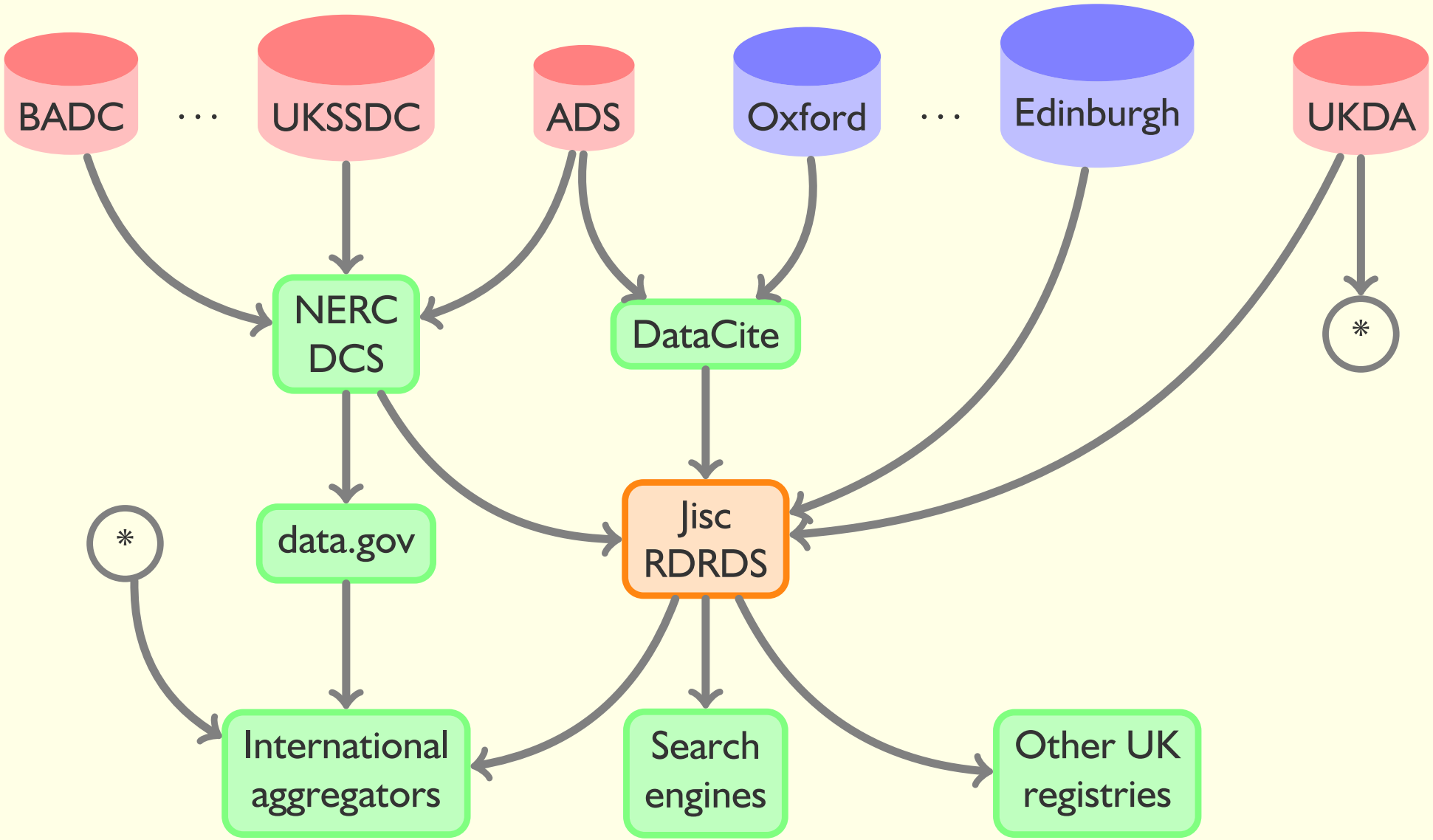
Discovery Service for UK Research Data

Scope and purpose

Discovery Service for UK Research Data



Where does RDRDS fit in?



Phase 1 pilot

- Registry based on ORCA (Research Data Australia)
- Participating data repositories:
  - 9 universities: Edinburgh, Glasgow, Hull, Lincoln, Leeds, Oxford, Oxford Brookes, St Andrews, Southampton
  - UKDA
  - Archaeology Data Service
  - 7 NERC Data Centres: BADC, BODC, EIDC, NEORC, NGDC, PDC, UKSSDC
- Crosswalks written *to* RIF-CS *from*
  - DataCite
  - DDI Codebook
  - EPrints (native + ReCollect)
  - MODS
  - OAI-PMH Dublin Core
  - UK Gemini 2

Search for Research Data

Browse by Subject Area

Browse by Map Coverage

Advanced Search

What's in the Research Data Registry and Discovery Service

Collections (49)
Research datasets or collections of research materials.

Parties (36)
Researchers or research organisations that create or maintain research datasets or collections.

Activities (0)
Projects or programs that create research datasets or collections.

Services (0)
Services that support the creation or use of research datasets or collections.

Spotlight on research data

http://rdrds.cloudapp.net/

Who contributes to the Research Data Registry and Discovery Service?

5 research organisations from around the UK contribute information to Research Data Australia.

See All

Share

Research Data Australia is an Internet-based discovery service designed to provide rich connections between data, projects, researchers and institutions, and promote visibility of Australian research data collections in search engines. Read more about us...

ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative.

Jisc

DCC

Quality levels for dataset records

Quality Level 2

- title
- description
- location (e.g. URL)
- IPR statement
- related party, e.g.
  - P.I./researcher
  - manager

Quality Level 3

- identifier
- citation information*
- subject
- date (e.g. of publication)
- spatial coverage
- temporal coverage
- related activity

* Such as ‘publisher’; other relevant fields are already mentioned.

Hydrographic data profiles collected by a conductivity-temperature-depth (CTD) sensor package during the Jan Mayen cruise JM4

Full Description

This dataset comprises 73 hydrographic data profiles, collected by a conductivity-temperature-depth (CTD) sensor package, in June 1994 from stations in the North East Norwegian Sea between 69 - 71 N, 15 - 19 E. A complete list of all data parameters are described by the SeaDataNet Parameter Discovery Vocabulary (PDV) keywords assigned in this metadata record. The data were collected by the University of Tromsø Norwegian College of Fishery Science as part of the Ocean Margin Exchange (OMEX) I project.

SHOW ALL DESCRIPTIONS

How to Cite this Collection

Citation (Metadata):

Tande, Kurt (2013,2013,2013,2010,2012): Hydrographic data profiles collected by a conductivity-temperature-depth (CTD) sensor package during the Jan Mayen cruise JM4. British Oceanographic Data Centre. Local: CSR9662CTDR00147. https://www.bodc.ac.uk/data/online_delivery/nodb/search/

Identifiers

Local: CSR9662CTDR00147

Additional Metadata

URI: http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetRecordById&ElementSetName=full&outputSchema=http://www.isotc211.org/2005/gmd&Id=b2535c18b9d9554fa24e25e50f3bf4a5

Dates

Issued: 2013-07-24 01:00

Created: 2010-02-03 01:00

Access

https://www.bodc.ac.uk/data/o...

Access rights

Usage restrictions are specified in the terms of the licence

Access rights

Data are freely available to all following agreement to the terms and conditions of the British Oceanographic Data Centre Data Licence. The licence terms and conditions are available via https://www.bodc.ac.uk/data/documents/nodb/267795/

Connections

People

Kurt Tande (PI)

Organisations & Groups

British Oceanographic Data Centre

Suggested Links

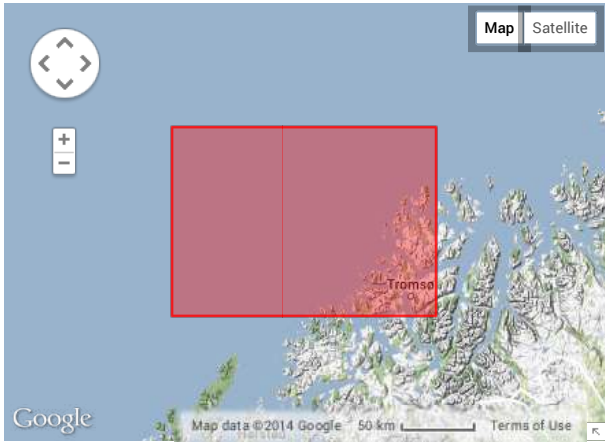
Internal Records

9 records with matching subjects

External Records

62 records from DataCite

Spatial Coverage:



text: Norwegian Sea

Temporal Coverage:

From 1994-06-13 01:00 to 1994-06-18 01:00

Subjects

Keywords

biota | oceans | Natural Environment Research Council Desi

Marine Environmental Data and Information Network | water col

upper epipelagic water column | water column | bathypelagic

mesopelagic water column | epipelagic water column

Coordinate reference systems | Elevation | Oceanographic g

Chlorophyll pigment concentrations in the water column | Densi

Salinity of the water column | Temperature of the water column

Vertical spatial coordinates

Phase 2

- Define a set of clear use cases and workflows.
- 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.

Metadata

Title

We found few problems here, except

- some records did not provide one;
- some titles were duplicated because they did not mention which subset was included;
- some might have to be redacted if they contain sensitive information.

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

Description/Abstract

- These were generally pretty good.
- There was variety in level of detail.

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

Dataset identifier

- Most places could provide a local ID.
- A handful supplied DOIs.
- One also supplied `<identifier type="citation">`.

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

Subject

- Most places could provide subject or topic terms.
- Some specified the scheme used.
- It was hard to transform these into our list of scheme identifiers.

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

URL of landing page

- Best case: derived from ID.
- Some provided multiple URLs: which to choose?

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

Creator (+ID)

- In one case, only provided within a citation.
- In another, provided as '`FirstName,LastName`' contrary to the specs.
- No IDs supplied: big trouble with duplication.

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

Release date

Various dates were supplied that could be used:

- published
- issued
- available

Should we choose? Take them all?

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

Rights information

Various types:

- access instructions
- access or usage restrictions
- licence statement
- licence URL

Not always easy to categorize them.

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

Spatial coverage

We only had a problem with `oai_dc` records – `dc:coverage` can contain

- coordinates of different sorts
- place or region names
- date or date range
- period name

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

Temporal coverage

- Not common, but usually consistent where provided.
- Only a problem in `dc:coverage`.

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Metadata

Publisher

- Tend to generate this from holding repository.
- Rarely provided explicitly.

- Title
- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Conclusions

- Ideal source format: one where there is only one right way of doing things!
- Need for identifiers all round:
  - datasets
  - people
  - organisations
  - subject vocabularies
  - subject terms
- People provide higher quality metadata if they see the effect it has.

Thank you for your attention

DCC Website: `http://www.dcc.ac.uk/`
Alex Ball: `http://alexball.me.uk/`

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