Metadata for Data Citation and Discovery

Alex Ball

DCC/UKOLN, University of Bath

6 July 2012

Except where otherwise stated, this work is licensed under Creative Commons Attribution 2.5 Scotland: http://creativecommons.org/licenses/by/2.5/scotland/
Outline

Citation, discovery and reuse

Data citation
  Motivation
  Metadata

Data discovery
  Motivation
  Metadata

Issues and challenges

Recommendations
Layers of metadata

Information Package

Content Information
- Data Object
- Representation Information

Packaging Information

Preservation Description Information
- Context Information
- Provenance Information
- Fixity Information
- Reference Information

Descriptive Information
Layers of metadata

Information Package

Content Information
- Data Object
- Representation Information

Preservation Description Information
- Context Information
- Provenance Information
- Fixity Information
- Reference Information

Packaging Information

Descriptive Information

1. Known item search
Layers of metadata

Information Package

Content Information
- Data Object
- Representation Information

Preservation Description Information
- Context Information
- Provenance Information
- Fixity Information
- Reference Information

1. Known item search
2. Speculative search

Descriptive Information

1. Known item search
2. Speculative search

DataCite Workshop 6 July 2012
Layers of metadata

- Content Information
- Representation Information
- Data Object

Information Package

Preservation Description Information
- Context Information
- Provenance Information
- Fixity Information
- Reference Information

1. Known item search
2. Speculative search
3. Reuse

DataCite Workshop 6 July 2012
What’s great about journal papers?

- Awareness raising
- Protection from plagiarism
- Verification of results
- Basis for future research
- Reward models
- Permanent access
What’s great about journal papers?

- Awareness raising
- Protection from plagiarism
- Verification of results
- Basis for future research
- Reward models
- Permanent access
Data citations provide:

- Visibility for data
- Protection from plagiarism
- Possibility for verification of results
- Data on which to base future research
- Possibility for reward models
- Access
Citation styles


Four data citation styles: which elements do they use?

Altman and King (2007): Dataverse

Lawrence et al. (2008): BADC

Green (2010): OECD

Starr and Gastl (2011): DataCite
Citation styles

Author

Altman and King (2007): Dataverse
  ➤ Sidney Verba.
  NORC [Producer];

Lawrence et al. (2008): BADC
  ➤ Iwi, A. and B. N. Lawrence

Green (2010): OECD
  ➤ OECD

Starr and Gastl (2011): DataCite
  ➤ Irino, T; Tada, R
Citation styles

Publication date

Altman and King (2007): Dataverse


NORC [Producer]:

Lawrence et al. (2008): BADC


Green (2010): OECD

OECD (2009),

(Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

Irino, T; Tada, R (2009):
Title

Altman and King (2007): Dataverse


Lawrence et al. (2008): BADC


Green (2010): OECD


Starr and Gastl (2011): DataCite

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797.
Altman and King (2007): Dataverse


Lawrence et al. (2008): BADC


Green (2010): OECD


Starr and Gastl (2011): DataCite

Feature

Altman and King (2007): Dataverse


Lawrence et al. (2008): BADC


Green (2010): OECD


Starr and Gastl (2011): DataCite

Altman and King (2007): Dataverse
- Sidney Verba. 1998. “U.S. and Russian Social and Political Participation Data,” NORC [Producer]; data set [Type (DC)]

Lawrence et al. (2008): BADC

Green (2010): OECD

Starr and Gastl (2011): DataCite
Publisher

Altman and King (2007): Dataverse


Lawrence et al. (2008): BADC


Green (2010): OECD


Starr and Gastl (2011): DataCite

Citation styles

Altman and King (2007): Dataverse

  NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2008): BADC


Green (2010): OECD

  (Accessed on 14 September 2009)

Starr and Gastl (2011): DataCite

Citation styles

Location

Altman and King (2007): Dataverse

NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2008): BADC

http://ndg.nerc.ac.uk/csml2/GridSeries] Version 1. BADC. urn:badc.nerc.ac.uk...coapec500yr [Available
from http://badc.nerc.ac.uk/data/coapec500yr].

Green (2010): OECD


Starr and Gastl (2011): DataCite

Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2. Geological
Institute, University of Tokyo. Dataset. doi:10.1594/PANGAEA.726855.
http://dx.doi.org/10.1594/PANGAEA.726855
Citation styles

Unique Numeric Fingerprint

Altman and King (2007): Dataverse

- UNF:3:ZNQRI14053UZq389x0Bffg== NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2008): BADC

- from http://badc.nerc.ac.uk/data/coapec500yr].

Green (2010): OECD


Starr and Gastl (2011): DataCite

- Institute, University of Tokyo. Dataset. doi:10.1594/PANGAEA.726855.
- http://dx.doi.org/10.1594/PANGAEA.726855
Key citation elements

- Author
- Publication date
- Title
- Location
Key citation elements

- Author
- Publication date
- Title
- Location (= identifier)
Motivation for a data discovery service

Needed to realise some benefits of data reuse:

- Maximum return on funder investment
- Reduced duplication of effort
- Reduced costs of data collection
- Broader scope of possible research
<table>
<thead>
<tr>
<th>Type</th>
<th>Dataset (5180)</th>
<th>Simulation</th>
<th>Metadata (3470)</th>
<th>Software (87)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td>CCSM (2752)</td>
<td>CMIP5 (1617)</td>
<td>GeoMIP (8)</td>
<td>NARCCAP (116)</td>
</tr>
<tr>
<td></td>
<td>NCL (47)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institute</td>
<td>NCAR (1461)</td>
<td>NSF-DOE-NCAR (356)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>CCSM3 (1807)</td>
<td>CCSM4 (1486)</td>
<td>CESM1-BGC (180)</td>
<td>CESM1-CAM5 (83)</td>
</tr>
<tr>
<td></td>
<td>CESM1-FASTCHEM (36)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment</td>
<td>1pcCO2 (52)</td>
<td>B04.10 (3)</td>
<td>B04.16 (4)</td>
<td>B04.19 (4)</td>
</tr>
<tr>
<td></td>
<td>B04.23 (4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product</td>
<td>forcing (4)</td>
<td>hist (326)</td>
<td>other (1)</td>
<td>output1 (1372)</td>
</tr>
<tr>
<td></td>
<td>output2 (245)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF Variable</td>
<td>10m wind speed (5)</td>
<td>2m air temperature (6)</td>
<td>2m reference specific humidity (6)</td>
<td>2m reference temperature (6)</td>
</tr>
<tr>
<td>Variable Name</td>
<td>ADVS (15)</td>
<td>ADVS_ISOP (8)</td>
<td>ADVS_SUBM (8)</td>
<td>ADVT (15)</td>
</tr>
<tr>
<td></td>
<td>ADVT_ISOP (8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensemble</td>
<td>r00p0 (136)</td>
<td>r10i1p1 (27)</td>
<td>r10i2p1 (54)</td>
<td>r11i1p1 (12)</td>
</tr>
<tr>
<td></td>
<td>r12i1p1 (12)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Domain</td>
<td>Atmosphere (794)</td>
<td>Earth system (138)</td>
<td>Ice (481)</td>
<td>Land (652)</td>
</tr>
<tr>
<td></td>
<td>Ocean (763)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Website powered by Science Gateway Framework (version 2.0.5-SNAPSHOT-20120702-010127)

Privacy Policy | Terms of Use
Details

**Title:** Chemistry of iron in freshwaters of Northwest England

**Abstract:** Chemical composition of freshwater samples from sites in Northern England. Measurements of pH, dissolved major ions (Na, Mg, K, Ca, Cl, NO3, SO4), dissolved organic carbon (DOC), dissolved Al, Fe(II) and total Fe, and measurements of Al, Fe(II) and total Fe on samples following dialysis. A view service identifying sample locations is also available.

**Temporal coverage:** These data span the period from Thursday 25 September 2003 to Thursday 26 February 2004 inclusive.

**Resource locator:**

Further Information - Preprint of peer-reviewed paper describing the results. Contains details of all analytical methods used.

**Miscellaneous**

Map View Service - This resource has a related map view service

Documents - Documents available to assist with re-use of this dataset.

**Data download** - Download this data

**Conditions for access and use constraints:** CCE must be acknowledged in all resultant publications

**Responsible party - Distributor**

Centre for Ecology & Hydrology

**Individual name:** Steve Loft

**Address:** Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford, OX10 8BB, United Kingdom

**Email:** enquiries(at)ceh(dot)ac(dot)uk
RAN XBT Temperature Profile Data - HMAS PERTH

This dataset contains temperature profile data recorded using expendable Bath thermograph probes (XBT) and a digital recording system. The dataset contains 50 observations obtained during RAN cruise 98004PE (HAMAS PERTH). Source file: 98004PE.xml

Coverage:

Time Period:
From 1998 To 1998

Subjects:
ANZSREC
STATISTICS
MATHEMATICAL PHYSICS
NUMERICAL AND
COMPUTATIONAL MATHEMATICS
EARTH SCIENCES
OTHER ENVIRONMENT

Keywords
Oceans | Ocean Temperature | Water Temperature

Identifiers:
Local: AODN:f311f2bb-3937-499e-8dcb-8f3d502b9709

http://services.ands.org.au/home/orca/rda/
Common data discovery metadata elements

Identification

- Dataset Name (15)
- Dataset Version (4)
- Dataset Date (13)
- Dataset Identifier (12)
- Metadata Scheme Name (7)
- Metadata Scheme Version (5)
- Metadata Record Date (10)
- Metadata Record Identifier (3)

Responsibility

- Project/Study/Series Name (9)
- Project/Study/Series Status (4)
- Rights/Restrictions (14)
- Agent (15)
- Agent Contact Details (11)
Common data discovery metadata elements

Archiving

- Location (15)
- File Format(s) (10)
- Storage Medium (6)
- Size (7)
- Data Quality Information (5)
- Data Preview (4)
- Dataset Language (7)
- Dataset Status (8)

Spatiotemporal Coverage

- Spatial Extent (12)
- Spatial Resolution (7)
- Temporal Extent (15)
- Temporal Resolution (5)
Common data discovery metadata elements

Topical Coverage and Derivation

- Dataset Type (12)
- Subject/Keywords (13)
- Abstract/Summary/Description (14)
- Parameters Used (6)
- Methodology/Instrumentation (8)
- Processing Steps (6)
- Related Datasets (11)
- Derived Publications (11)

http://www.ukoln.ac.uk/projects/sdapss/
Most common data discovery metadata elements

- Dataset Name (15)
- Dataset Date (13)
- Dataset Identifier (12)
- Metadata Record Date (10)
- Rights/Restrictions (14)
- Agent (15)
- Agent Contact Details (11)
- Location (15)
- File Format(s) (10)
- Spatial Extent (12)
- TemporalExtent (15)
- Dataset Type (12)
- Subject/Keywords (13)
- Abstract/Summary/Description (14)
- Related Datasets (11)
- Derived Publications (11)
Most common data discovery metadata elements

- Dataset Name (15)
- Dataset Date (13)
- Dataset Identifier (12)
- Metadata Record Date (10)
- Rights/Restrictions (14)
- Agent (15)
- Agent Contact Details (11)
- Location (15)
- File Format(s) (10)
- Spatial Extent (12)
- Temporal Extent (15)
- Dataset Type (12)
- Subject/Keywords (13)
- Abstract/Summary/Description (14)
- Related Datasets (11)
- Derived Publications (11)
Most common data discovery metadata elements

- Dataset Name (15)
- Dataset Date (13)
- Dataset Identifier (12)
- Metadata Record Date (10)
- Rights/Restrictions (14)
- Agent (15)
- Agent Contact Details (11)
- Location (15)
- File Format(s) (10)
- Spatial Extent (12)
- Temporal Extent (15)
- Dataset Type (12)
- Subject/Keywords (13)
- Abstract/Summary/Description (14)
- Related Datasets (11)
- Derived Publications (11)
Most common data discovery metadata elements

- Dataset Name (15)
- Dataset Date (13)
- Dataset Identifier (12)
- Metadata Record Date (10)
- Rights/Restrictions (14)
- Agent\(^1\) (15)
- Agent Contact Details (11)
- Location\(^2\) (15)
- File Format(s) (10)
- Spatial Extent (12)
- Temporal Extent (15)
- Dataset Type (12)
- Subject/Keywords (13)
- Abstract/Summary/Description (14)
- Related Datasets (11)
- Derived Publications (11)

1. DataCite uses Creator, Publisher, Contributor
2. DataCite holds this separately
3. DataCite also has Dataset Version (4), Dataset Language (7), Size (7)
### Attributing datasets to many contributors

DataCite Workshop | 6 July 2012

[![Attributing datasets to many contributors](image)]
Dynamic datasets

Two types:

- Revised datasets
- Expanding datasets
Dynamic datasets

Three strategies:

1. Differentiate versions by access date rather than ID
   - A

2. Take time slices
   - A
   - B
   - C

3. Take snapshots
   - A
   - B
   - C
Recommendations: metadata for citation

- **Author**
  - Record roles, identifiers and contact details as well as names.

- **Publication date**

- **Title**
  - It helps to avoid confusion if this is different from the article title.

- **Location/Identifier**
  - Express identifiers in location (http) form if possible.
  - Use DOIs for ‘published’ data and Handles, ARKs, PURLs, etc. for other data.
  - Use different DOIs for different versions.
  - Location given should be a catalogue record/landing page for the data.
  - Giving the **Publisher** or **Host Archive** as well provides some recourse if the link breaks.
Recommendations: metadata for discovery

In addition to the citation metadata, these should also be recorded where they apply:

- Contributors
- Abstract/Summary/Description
- Subject/Keywords
- Rights/Restrictions
- Spatial Coverage
- Temporal Coverage
- Derived Publications
- Related Datasets
- Resource Type
- File Format(s)
- Important Dates
  - Creation, Submission, Acceptance, Use By…
- Language*
- Version*
- Size*
- Metadata Record Date
Thank you for your attention

DCC Website: http://www.dcc.ac.uk/
Alex Ball: http://www.ukoln.ac.uk/ukoln/staff/a.ball/

8th International Digital Curation Conference
“Infrastructure, Intelligence, Innovation: driving the Data Science agenda”
14–16 January 2013, Amsterdam
http://www.dcc.ac.uk/events/idcc13