A Data Citation Roadmap for Scholarly Data Repositories

Tim Clark (Harvard Medical School & Massachusetts General Hospital)
Martin Fenner (DataCite)
Mercè Crosas (Institute for Quantitative Social Science, Harvard University)

DataCite Webinar, February 23, 2017
Background

• NIH, NAS, other science policy makers very concerned about scientific reproducibility & robustness of results ¹.

• Significant science policy studies recommend archiving & direct citation of primary data in research articles ², ³, ⁴.

• NIH Big Data to Knowledge (BD2K) Program: “Facilitate broad use of biomedical digital assets by making them discoverable, accessible and citable.” (NIH 2015) ⁵

• Technology and many recommendations in place ⁶, ⁷.

• NIH-funded BD2K program bioCADDIE for data discovery ⁸.
Some reasons to cite data

1. Transparency & Validation => better science
   Reproducibility & Robustness

2. Big Data meta-analyses => re-use & discovery
   Extract new knowledge

3. Radically Improve Biomedical Translation => cure diseases
Joint Declaration of Data Citation Principles

JDDCP endorsed by over 100 scholarly organizations
Achieving human and machine accessibility of cited data in scholarly publications

Joan Starr¹, Eleni Castro², Mercè Crosas², Michel Dumontier³, Robert R. Downs⁴, Ruth Duerr⁵, Laurel L. Haak⁶, Melissa Haendel⁷, Ivan Herman⁸, Simon Hodson⁹, Joe Hourclé¹⁰, John Ernest Kratz¹, Jennifer Lin¹¹, Lars Holm Nielsen¹², Amy Nurnberger¹³, Stefan Proell¹⁴, Andreas Rauber¹⁵, Simone Sacchi¹³, Arthur Smith¹⁶, Mike Taylor¹⁷, and Tim Clark¹⁸

¹California Digital Library, Oakland CA US
²Harvard University, Institute of Quantitative Social Sciences, Cambridge MA US
³University, Palisades, New York US
⁴National Snow and Ice Data Center, Boulder CO US
⁵ORCID, Inc., Bethesda MD US
⁶Oregon Health and Science University, Portland OR US
⁷W3C/CWI, Amsterdam, the Netherlands
⁸CODATA (ICSU Committee on Data for Science and Technology), Paris FR
⁹Solar Data Analysis Center, NASA Goddard Space Flight Center, Greenbelt MD US
¹⁰Public Library of Science, San Francisco CA US
¹¹European Organization for Nuclear Research (CERN), Geneva CH
¹²Columbia University Libraries/Information Services, New York NY US
¹³SBA Research, Vienna AT
¹⁴Institute of Software Technology and Interactive Systems, Vienna University of Technology / TU Wien, AT
¹⁵American Physical Society, Ridge NY US
¹⁶Elsevier, Oxford UK
¹⁷Harvard Medical School, Boston MA US

Direct deposition and citation of primary research data

http://doi.org/10.7717/peerj-cs.1
Data Citation Implementation Pilot

2016
Participants

And you!

Springer

The Dataverse Project

EMBL-EBI
Data Citation Generic Example

example of a data citation as it would appear in a reference list

**Principle 2: Credit and Attribution**
(e.g. authors, repositories or other distributors and contributors)

**Principle 4: Unique Identifier**
(e.g. DOI, Handle.). **Principle 5, 6 Access, Persistence:**
A persistent link to a landing page with metadata and access information

**Author(s), Year, Dataset Title, Data Repository or Archive, [Accession], Global Persistent Identifier, version or subset**

**Principle 7: Version and granularity**
(e.g. a version number or a query to a subset) In addition, access to versions or subsets should be available from the landing page.

*Note that the format is not intended to be defined with this example, as formats will vary across publishers and communities [Principle 8: Interoperability and flexibility].
Role-based Participants

• **Publishers**

• **Data Repositories**
  • EMBL, Dataverse, Dryad, Figshare, Google, etc.

• **Informaticians** (NIH BD2K, EBI, CDL, etc.)

• … & Authors (YOU)
Publishers Roadmap Development

- Leads: Amye Kenall & Helena Cousijn
- Workshop July 22 @ SpringerNature London campus, partially funded by NPG.
- Continuing work via Telcons.
Data citation at Springer Nature journals – key events

• 1998 – : Accession codes required for various data types at Nature journals and marked up in articles (= data referencing rather than formal citation)

• 2012: Data citation included in BMC style guide for all its journals
  
  https://blogs.biomedcentral.com/bmcblog/2012/01/19/citing-and-linking-data-to-publications-more-journals-more-examples-more-impact/

Publishers are taking data citation seriously

• 2014: NPG Signatory of Joint Declaration of Data Citation Principles
  
  http://blogs.nature.com/scientificdata/2014/03/24/endorsing-the-joint-declaration-of-data-citation-principles/

• 2014: Launch of Scientific Data
  
  • Data citation mandated for every article
  
  • Uses JATS 1.0 with data citations list specifically tagged

• 2016: Data citation policy piloted at Nature journals
  
  • Strongly encourages datasets with DOIs to be included in reference lists

• 2016: Springer Nature wide project to support data citation in all journals’ policies

adapted with permission from a talk by Ian Hrynaszkwiewicz, July 2016
Publisher’s Roadmap
Approach & Status

- **Roadmap** based on experiences of early adopter publishers.
  - Examples, real situations, recommended approaches.
  - Complete end-to-end publishing workflow.


- Cousijn et al. 2017 *bioRxiv* [https://doi.org/10.1101/100784](https://doi.org/10.1101/100784)
Elsevier Implements Data Citation Standards to Encourage and Reward Authors for Sharing Research Data

Amsterdam, November 30, 2016

Elsevier, a world-leading provider of scientific, technical and medical information products and services, today announced that it has implemented the FORCE11 Joint Declaration of Data Citation Principles for over 1,800 journals. This means that authors publishing with Elsevier are now able to cite the research data underlying their article, contributing to attribution and encouraging research data sharing with research articles.

The FORCE11 data citation principles were launched in 2014 with the aim to make research data an integral part of the scholarly record. The principles recognized that a critical driver for increasing the availability of research data was to ensure authors receive credit for sharing through proper citation of research data. Elsevier was involved in drafting these principles and, along with many other publishers, data repositories and research institutions, endorsed them as an industry standard. Now, after working closely with other publishers within the Data Citation Implementation Pilot, Elsevier has incorporated them in its production and publication workflow in order to recognize and process data citations. Combined with new author guidance and education, this will encourage and reward researchers for sharing their research data.
Repository Metadata Expert Group
Repositories Roadmap Approach & Status

- Roadmap rev 1 specifies core data citation metadata.
  - What must be on landing pages & how it is made machine-readable.
- Preprint published December 28, 2016
  - >1K downloads in 42 days.
- Fenner et al. 2016 *bioRXiv* [https://doi.org/10.1101/097196](https://doi.org/10.1101/097196).
  - Rev 2 being developed based on community feedback, including schema.org initiative.
Article - Landing Page - Data

Bibliographic Repository

Article 1
Finding 2
Data reference 3

Data Repository

Landing Page 2
Human readable metadata
Machine readable metadata
article PID
citation metadata
data PID

Data 3

Bibliographic Repository - Data Repository - Landing Page - Data

1 1 0 0 1 0 0 0 1 1 0 1 0
1 0 1 1 0 1 1 1 0 0 1 1
1 1 1 0 0 0 0 1 1 1 1 1
1 1 1 1 0 1 0 1 0 0 1 0
0 1 0 0 1 1 1 1 0 1 0 0 1
0 1 1 0 1 1 1 0 0 1 1 1 0
1 1 1 0 0 1 0 0 1 1 1 0

MeSH

PubMed

DataMed

DATS
Repositories: Required

1. All datasets intended for citation must have a globally unique persistent identifier that can be expressed as unambiguous URL.

2. Persistent identifiers for datasets must support multiple levels of granularity, where appropriate.

3. This persistent identifier expressed as URL must resolve to a landing page specific for that dataset.

4. The persistent identifier must be embedded in the landing page in machine-readable format.

5. The repository must provide documentation and support for data citation.
Globally Unique Persistent Identifier

- **Persistent method for identification**: Metadata must persist even beyond the data it describes.

- **Machine actionable**: PID resolvable as an HTTP URI.

- **Globally Unique**: Must use a prefix if ID only unique within a database.

- **Widely used by a community**: For example, in life sciences accession numbers (not DOIs) are widely used.
Multiple Levels of Granularity

- Support citation of a specific version, as well as citation of unspecified version

In some cases, data is uniquely identified as a collection of many items (example in next slide)
Image collection 10.18116/C6H02X

UMass/CANDI Image Attribution Framework, 2016

DOI 10.18116/C6H02X

DataCite XML

If you are citing this data because of one of the references below, please cite the reference of interest. If you are citing this data in its own right (independent of any of the references below), we suggest the following citation (APA): Breeze, JL, Caplan, D, Caviness, VS, Frazier, JA, Giuliano, AJ, Haselgrove, C, ... Zablotsky, B. (2016). Image collection 10.18116/C6H02X. UMass/CANDI Image Attribution Framework. http://dx.doi.org/10.18116/C6H02X.

Refine/download

<table>
<thead>
<tr>
<th>Description</th>
<th>PubMed ID</th>
<th>Publication DOI</th>
<th>Funder</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>This collection contains all images (structural scans and segmentations) for female subjects age 10 and above from the Internet Brain Segmentation Repository and CANDI Share Schizophrenia Bulletin 2008 data sets. It was created as a demonstration collection.</td>
<td></td>
<td></td>
<td>NIMH</td>
<td>Honor, Leah Haselgrove, Christian Frazier, Jean A Kennedy, David N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source projects</th>
<th>10.18116/C6WC71</th>
<th>10.18116/C6159Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source images</td>
<td>10.18116/C67P43</td>
<td>10.18116/C63W25</td>
</tr>
<tr>
<td></td>
<td>10.18116/C6059N</td>
<td>10.18116/C6VC7Q</td>
</tr>
<tr>
<td></td>
<td>10.18116/C66P4S</td>
<td>10.18116/C6301T</td>
</tr>
<tr>
<td></td>
<td>10.18116/C6759R</td>
<td></td>
</tr>
</tbody>
</table>

https://doi.org/10.18116/C6H02X
Persistent Identifier resolves to Landing Page

- Using HTTP redirection makes it easier to maintain a stable URL for the persistent identifier

- Identifiers.org, DOIs, handles and ARKs all use redirection

- Expectation is that the persistent identifier resolves to a human-readable page with more information

- (optional) Use content negotiation to resolve the persistent identifier URL to machine-readable metadata, or to the content itself
Persistent Identifier embedded in the Landing Page

Human Readable

Machine Readable

Example schema.org/JSON-LD

```json
<application type="application/ld+json">
  {
    "@id": "https://doi.org/10.5061/dryad.q447c/3"
  }
</application>
```

Example HTML meta tags

```html
<meta name="DC.identifier" content="https://doi.org/10.5061/dryad.q447c/3">
```
Documentation and Support

• The repository must provide documentation about how data should be cited, how metadata can be obtained, and who to contact for more information.

• The DCIP FAQ Expert Group has generated example documentation for data repositories, which will be provided on a dedicated website.
Repositories: Recommended

6. The landing page should include metadata required for citation, and ideally also metadata helping with discovery, in human-readable and machine-readable format.

7. The machine-readable metadata should use schema.org markup in JSON-LD format.

8. Metadata should be made available via HTML meta tags to facilitate use by reference managers.
## Citation Metadata

<table>
<thead>
<tr>
<th>Citation Metadata</th>
<th>Dublin Core(^a)</th>
<th>Schema.org(^b)</th>
<th>DataCite(^c)</th>
<th>DATS(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset Identifier</td>
<td>identifier</td>
<td>@id(^*)</td>
<td>identifier</td>
<td>identifier</td>
</tr>
<tr>
<td>Title</td>
<td>title</td>
<td>name</td>
<td>title</td>
<td>title</td>
</tr>
<tr>
<td>Creator(^**)</td>
<td>creator</td>
<td>author</td>
<td>creator</td>
<td>creator</td>
</tr>
<tr>
<td>Data repository or archive</td>
<td>publisher</td>
<td>publisher</td>
<td>publisher</td>
<td>publisher</td>
</tr>
<tr>
<td>Publication Date</td>
<td>date</td>
<td>datePublished</td>
<td>publicationYear</td>
<td>date</td>
</tr>
<tr>
<td>Version</td>
<td>not available</td>
<td>version</td>
<td>version</td>
<td>version</td>
</tr>
<tr>
<td>Type</td>
<td>type</td>
<td>type</td>
<td>resourceTypeGeneral</td>
<td>type</td>
</tr>
</tbody>
</table>
Metadata on Landing Pages: For Humans

<table>
<thead>
<tr>
<th>Dataset Persistent ID</th>
<th>hdl:1902.1/00341</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication Date</td>
<td>2010-04-01</td>
</tr>
<tr>
<td>Title</td>
<td>Careers, Marriage, Identity, and Feminism: Women's Life-Choices in the Seventies, 1975</td>
</tr>
<tr>
<td>Other ID</td>
<td>00341</td>
</tr>
<tr>
<td>Author</td>
<td>Charlotte Weissberg (Deceased)</td>
</tr>
<tr>
<td>Contact</td>
<td>Use email button above to contact.</td>
</tr>
<tr>
<td>Description</td>
<td>This study examined the personalities of a group of college women and their interactions with social institutions. The focus was on the relationship between the changes in values and ideas brought about by the women's movement and the personal development of young women.</td>
</tr>
</tbody>
</table>
Metadata on Landing Pages: schema.org

Schema.org is community activity to promote structured data on the internet, started in 2011 by Google, Microsoft, Yahoo, and Yandex.

Schema.org can be displayed as microdata or RDFa embedded in HTML, or via JSON-LD. JSON-LD is the preferred format for data citation metadata.

Citation metadata are fully supported by schema.org (see earlier citation metadata table), several groups are extending support for more specialized metadata, including http://bioschemas.org in the life sciences.

DataCite has released a command-line tool (https://github.com/datacite/bolognese) to automatically generate schema.org/JSON-LD for DataCite and Crossref DOIs, making it easier for data centers to integrate schema.org in landing pages.
Schema.org Example

{
    "@context": "http://schema.org",
    "@type": "Dataset",
    "@id": "https://doi.org/10.18116/c6h02x",
    "additionalType": "Imaging Data",
    "name": "Image collection 10.18116/C6H02X",
    "alternateName": "http://iaf.virtualbrain.org/search/reconstitute/2ccda04d",
    "author": [{
        "@type": "Person",
        "givenName": "JL",
        "familyName": "Breeze"
    }, ...]
Metadata on Landing Pages: HTML Meta Tags

<meta name="DC.identifier" content="doi:10.1594/PANGAEA.727206" scheme="DCTERMS.URI" />
<meta name="DC.title" content="Landings of European lobster (Homarus gammarus) and edible crab (Cancer pagurus) from 1615 to 2009, Helgoland, North Sea" />
<meta name="DC.creator" content="Schmalenbach, Isabel" />
<meta name="DC.creator" content="Mehrtens, Folke" />
<meta name="DC.creator" content="Janke, Michael" />
<meta name="DC.creator" content="Buchholz, Friedrich" />
<meta name="DC.publisher" content="PANGAEA" />
<meta name="DC.date" content="2011-01-28" scheme="DCTERMS.W3CDTF" />
<meta name="DC.type" content="Dataset" />
9. **Content negotiation** for schema.org/JSON-LD and other content types may be supported so that the persistent identifier expressed as URL resolves directly to machine-readable metadata.

10. **HTTP link headers** may be supported to advertise content negotiation options.

11. Metadata may be made available for download in Bibtex or other standard bibliographic format.
Content Negotiation for Machine Readable Metadata

Example Image Attribution Framework (IAF)

Examples DataCite
curl -LH "Accept: application/ld+json" http://doi.org/10.5061/DRYAD.8290N
Example

curl -I https://search.datacite.org/works/10.5061/dryad.q447c/3

HTTP/1.1 200 OK
Content-Type: text/html; charset=utf-8
Status: 200 OK
Link: <https://doi.org/10.5061/dryad.q447c/3> ; rel="identifier",
    <https://doi.org/10.5061/dryad.q447c/3> ; rel="describedby"
    type="application/vnd.datacite.datacite+xml",
    <https://doi.org/10.5061/dryad.q447c/3> ; rel="describedby"
    type="application/ld+json",
    <https://doi.org/10.5061/dryad.q447c/3> ; rel="describedby"
    type="application/vnd.citationstyles.csl+json",
    <https://doi.org/10.5061/dryad.q447c/3> ; rel="describedby"
    type="application/x-bibtex"
Metadata in Standard Bibliographic Format

Example BibTex

@data{25240_2014,
    author = {Figueiredo, Dalson and Rocha, Enivado and Paranhos, Ranulfo and Alexandre, José},
    publisher = {Harvard Dataverse},
    title = {How can soccer improve statistical learning?},
    year = {2014},
    doi = {10.7910/DVN/25240},
    url = {https://doi.org/10.7910/DVN/25240}
}

Example RIS

TY - DATAT1 - How can soccer improve statistical learning?
A1 - Figueiredo, Dalson
A1 - Rocha, Enivaldo
A1 - Paranhos, Ranulfo
A1 - Alexandre, José
Y1 - 2014
DO - 10.7910/DVN/25240
UR - https://doi.org/10.7910/DVN/25240
ER -
Conclusions

- We need to systematically cite data for improved scientific transparency, reproducibility, robustness.

- Persistent discoverable data archives with cited data will enhance capability for validation & re-use.

- DCIP promotes data citation in journals, repositories, and identifier / metadata services at scale.

- **Rev 1 Roadmaps and specs released in bioRXiv**

- Continuing outreach, documentation and discussion.
DCIP Executive

- Tim Clark, Harvard Medical School & MGH (co-Chair)
- Maryann Martone, Hypothesis & UCSD (co-Chair)
- Carole Goble, University of Manchester & ELIXIR
- Jeffrey Grethe, UCSD & bioCADDIE
- Jo McEntyre, EMBL-EBI & ELIXIR
- Joan Starr, California Digital Library
- Martin Fenner, DataCite
- Simon Hodson, CODATA
- Chun-Nan Hsu, UCSD