

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 discovery⁸

Some reasons to cite data

1

- Transparency & Validation => better science
- Reproducibility & Robustness

2

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

3

- Radically Improve Biomedical Translation
=> cure diseases

1. Importance

Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications^[1].

2. Credit and Attribution

Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data^[2].

3. Evidence

In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited^[3].

4. Unique Identification

A data citation
community

Joint Declaration of Data Citation Principles

5. Access

Data citations
are necessary

JDDCP endorsed by over 100 scholarly organizations

materials, as

6. Persistence

Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe^[6].

7. Specificity and Verifiability

Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited^[7].

8. Interoperability and Flexibility

Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities^[8].

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

Direct deposition and citation of primary research data

³University, Fairport, New York US

⁵National Snow and Ice Data Center, Boulder CO US

⁶ORCID, Inc., Bethesda MD US

⁷Oregon Health and Science University, Portland OR US

⁸W3C/CWI, Amsterdam, the Netherlands

⁹CODATA (ICSU Committee on Data for Science and Technology), Paris FR

¹⁰Solar Data Analysis Center, NASA Goddard Space Flight Center, Greenbelt MD US

¹¹Public Library of Science, San Francisco CA US

¹²European Organization for Nuclear Research (CERN), Geneva CH

¹³Columbia University Libraries/Information Services, New York NY US

¹⁴SBA Research, Vienna AT

¹⁵Institute of Software Technology and Interactive Systems, Vienna University of Technology / TU Wien, AT

¹⁶American Physical Society, Ridge NY US

¹⁷Elsevier, Oxford UK

¹⁸Harvard Medical School, Boston MA US

Data Citation Implementation Pilot

The screenshot shows the BioCADDIE website homepage. At the top left is the BioCADDIE logo with the tagline "biomedical and healthCare Data Discovery Index Ecosystem". To the right is a login form with fields for "E-mail or username *" and "Password *", a "Log in" button, and links for "Create new account" and "Request new password". Below the header is a navigation menu with items: HOME, ABOUT, GROUPS, PARTICIPATE, NEWS, EVENTS, RESOURCES, RELATED LINKS, and CONTACT. The main banner features a server room background with the BioCADDIE logo. Below the banner are three content blocks: "ANNOUNCEMENTS" with three items, "ACCESS DATAMED" with a DataMed logo on a monitor, and "METADATA STANDARDS" with a diagram of a DataMed database and its fields (ID, Org, Repo, PMID, ISSN, Date).

bioCADDIE biomedical and healthCare Data Discovery Index Ecosystem

E-mail or username * Password *
Create new account Request new password

HOME ABOUT GROUPS PARTICIPATE NEWS EVENTS RESOURCES RELATED LINKS CONTACT

ANNOUNCEMENTS

- The DataMed's DATS Model Annotated With Schema.org - Webinar
- BioCADDIE Repository Workshop June 23, 2016
- Sign Up For Access To DataMed (BioCADDIE Prototype)
- Pilot Project On Harvester Announcement

ACCESS DATAMED

DataMed

METADATA STANDARDS

Pin it

Diagram illustrating DataMed metadata standards:

- ID
- Org
- Repo
- PMID
- ISSN
- Date

Participants



*And
you!*



Springer



eLIFE



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**
 - Elsevier, Springer Nature, PLOS, eLife, Wiley, Frontiers, etc. ...
- **Data Repositories**
 - EMBL, Dataverse, Dryad, Figshare, Google, etc.
- **Informaticians** (NIH BD2K, EBI, CDL, etc.)
- ... & Authors (YOU)

Publishers Roadmap Development

GROUP
LEADER

Helena Cousijn

Elsevier



- Leads: Amye Kenall & Helena Cousijn
- Participants: Elsevier, SpringerNature, eLife, PLoS, Frontiers, Wiley, et al.

GROUP
LEADER

Amye Kenall

SpringerNature



- 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 Hrynaszkiewicz, 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.
- Preprint published January 19, 2017.
- Cousijn et al. 2017 *bioRxiv* <https://doi.org/10.1101/100784> .

Publishers Roadmap adoption: 1,800 Elsevier journals as of Nov. 30, 2016

[Elsevier](#) > [About](#) > [Press releases](#) > [Science & technology](#) > [Elsevier Implements ...](#)

Elsevier Implements Data Citation Standards to Encourage and Reward Authors for Sharing Research Data

Share this:

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 1800 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

Leads: Martin Fenner (DataCite), Merce Crosas (Dataverse)

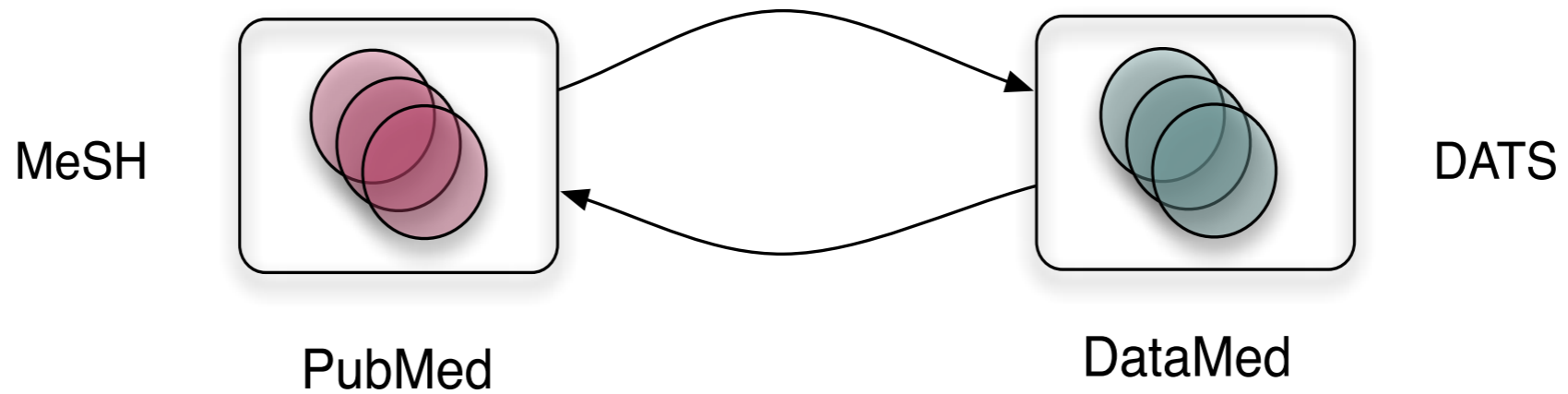
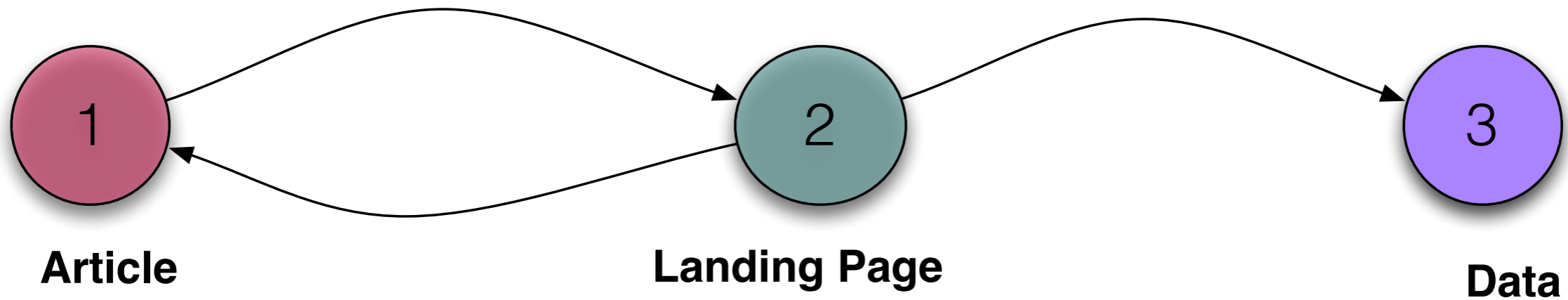


Repository Metadata Expert Group

Repositories Roadmap Approach & Status

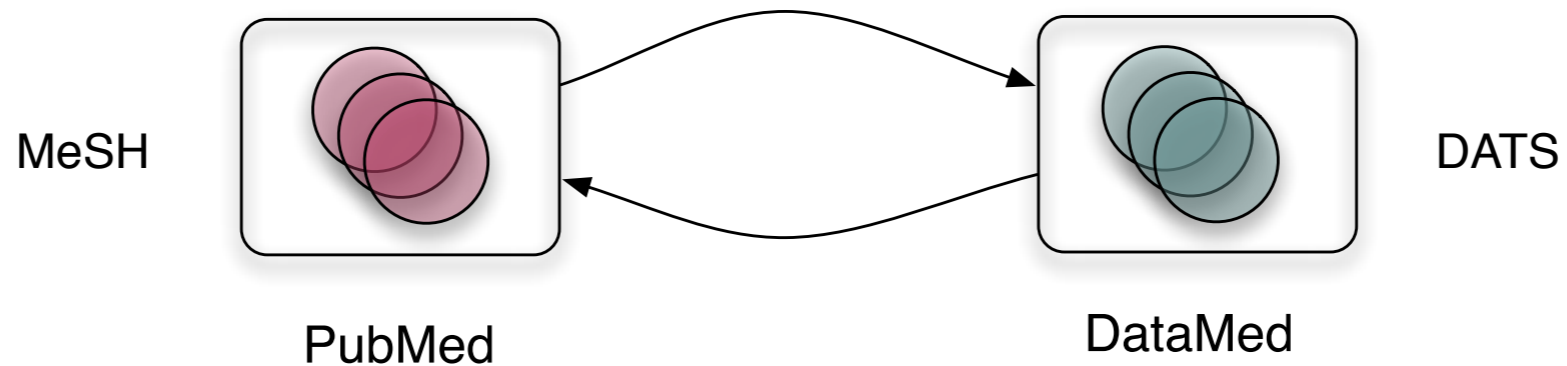
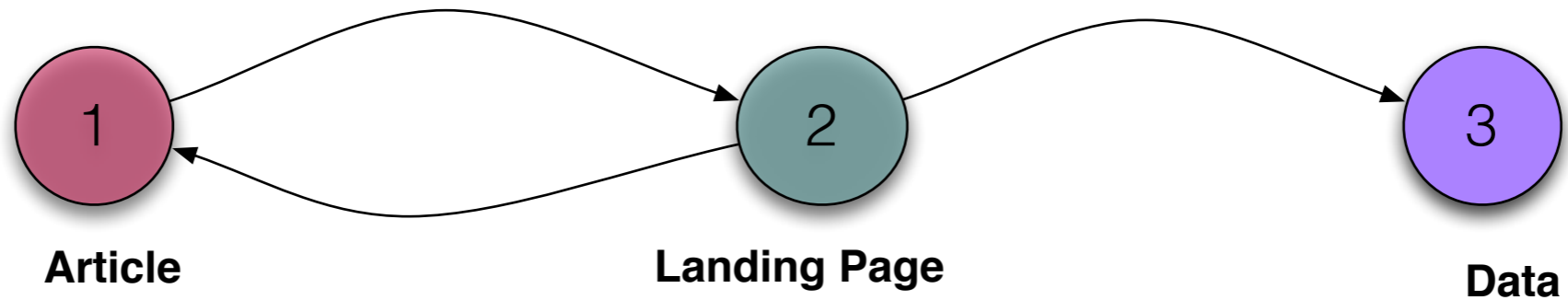
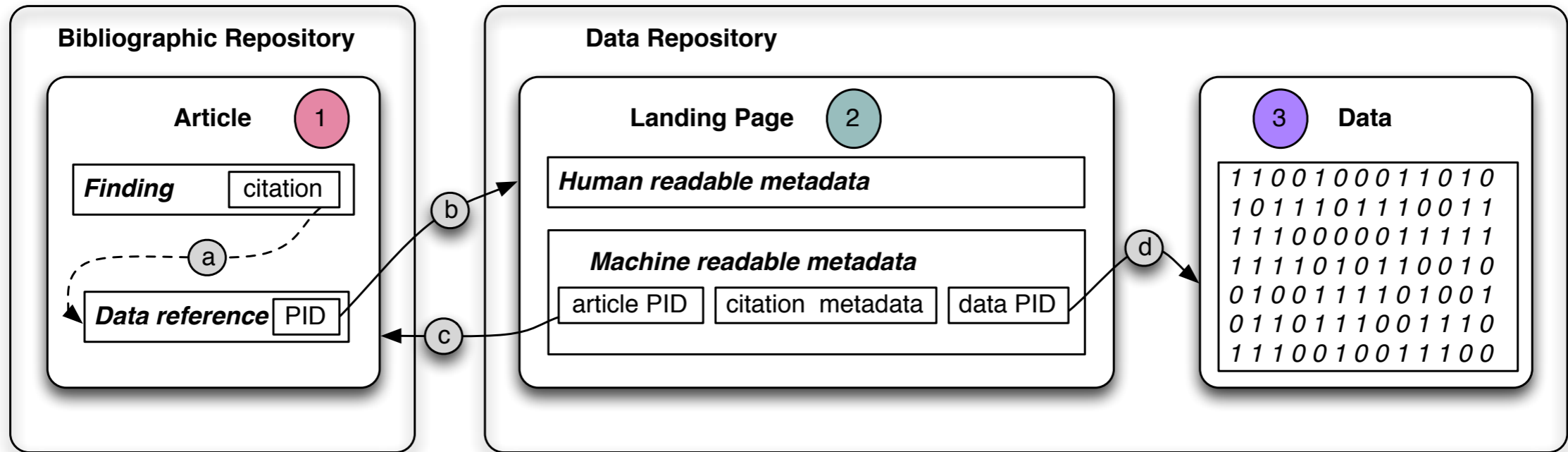
- Roadmap rev 1 specifies core data citation metadata.
 - What must be on landing pages & how it is made machine-readable.
- Preprint published December 28, 2016
 - >1K downloads in 42 days.
- Fenner et al. 2016 *bioRxiv* <https://doi.org/10.1101/097196>.
 - Rev 2 being developed based on community feedback,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

Article - Landing Page - Data



Discovery Indexes

Article - Landing Page - Data



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

Charlotte Weissberg (Deceased), 2010, "Careers, Marriage, Identity, and Feminism: Women's Life-Choices in the Seventies, 1975", [hdl:1902.1/00341](https://doi.org/10.191/00341), Harvard Dataverse, V5

 Cite Dataset ▾

 Learn about Data Citation Standards.

- 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](#)

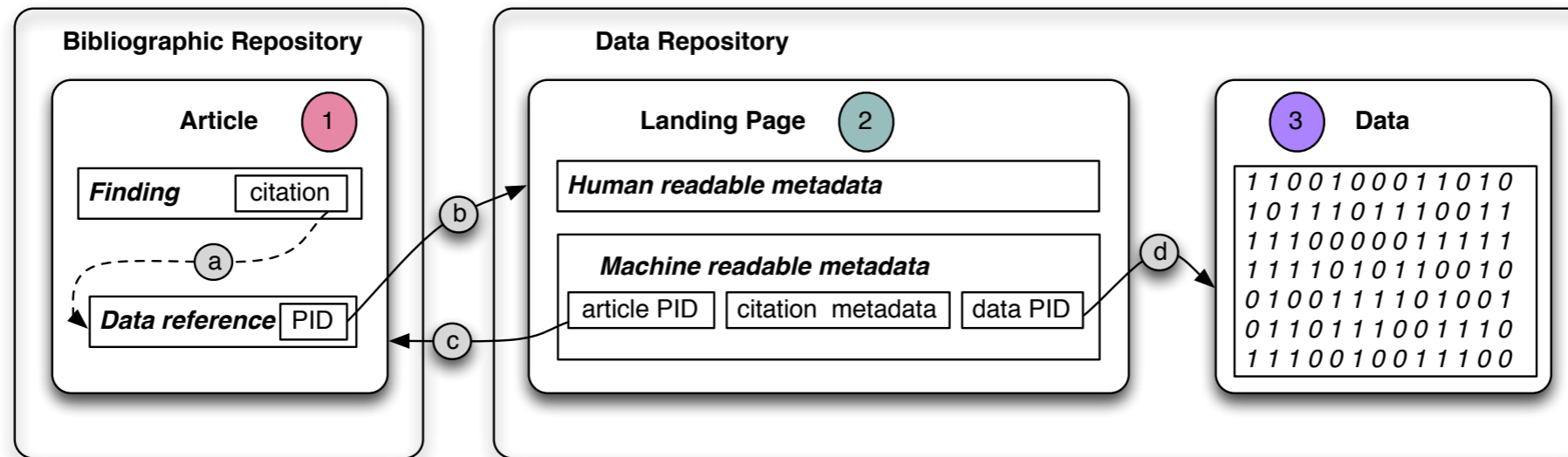
Description	PubMed ID	Publication DOI	Funder	Authors
This collection contains all images (structural scans and segmentations) for female subjects age 10 and above from the Internet Brain Segmentation Repository and CANDI Share Schizophrenia Bulletin 2008 data sets. It was created as a demonstration collection.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

Source projects	10.18116/C6WC71 10.18116/C6159Z
Source images	10.18116/C67P43 10.18116/C63W25 10.18116/C6059N 10.18116/C6VC7Q 10.18116/C66P4S 10.18116/C6301T 10.18116/C6759R

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

Cite this Dataset

Bilokapic, S; Schwartz, TU. 2015. "X-Ray Diffraction data for: Nup37-Nup120 full-length complex from Schizosaccharomyces pombe. PDB Code 4FHN", SBGrid Data Bank, V1,

<http://dx.doi.org/10.15785/SBGRID/179>.

[Download Citation](#)

Machine Readable

Example schema.org/JSON-LD

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

Example HTML meta tags

```
<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

Citation Metadata	Dublin Core ^a	Schema.org ^b	DataCite ^c	DATS ^d
Dataset Identifier	identifier	@id*	identifier	identifier
Title	title	name	title	title
Creator**	creator	author	creator	creator
Data repository or archive	publisher	publisher	publisher	publisher
Publication Date	date	datePublished	publicationYear	date
Version	<i>not available</i>	version	version	version
Type	type	type	resourceTypeGeneral	type


Metadata on Landing Pages: For Humans

[Files](#)


[Metadata](#)

[Terms](#)

[Versions](#)

 [Export Metadata](#) ▾

Citation Metadata

Dataset Persistent ID	hdl:1902.1/00341
Publication Date	2010-04-01
Title	Careers, Marriage, Identity, and Feminism: Women's Life-Choices in the Seventies, 1975
Other ID	00341
Author	Charlotte Weissberg (Deceased)
Contact	 Use email button above to contact.
Description	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.

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" />
```

Recommendations: Optional

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)

```
curl -H "Accept: application/xml"  
http://iaf.virtualbrain.org/lp/10.18116/C6WC71
```

Examples DataCite

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

```
curl -LH "Accept: application/vnd.citationstyles.csl+json"  
http://doi.org/10.5061/DRYAD.8290N
```

HTTP Link Headers

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, Enivaldo 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