Force 11’s Data Citation Activities: A Quick Summary

Tim Clark
Harvard Medical School & Massachusetts General Hospital

Joan Starr
California Digital Library

National Academy of Sciences, Washington DC

July 12, 2016
Joint Declaration of Data Citation Principles

**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 is an identifier of a data resource within a community.[4]

**5. Access**
Data citations are necessary to access data.[5]

**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]

[1] Joint Declaration of Data Citation Principles (JDDCP) endorsed by over 100 scholarly organizations

http://force11.org/datacitation
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].
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 NY 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
Recommendations for accessibility

1. Unique identification
2. Landing Pages
3. Persistence guarantees for repositories
What should endorsing the JDDCP mean?

1. Archives and repositories
2. Registries
3. Researchers
4. Funding agencies
5. Scholarly Societies
6. Academic institutions
Data Citation Implementation Pilot
Pilot Strategic Objectives

a. Provide coordination & guidance for early adopters.
b. Help establish benchmark implementations.
c. Focus on archiving and citing primary research data.
d. Provide report on lessons learned to the community.
e. Make cited data discoverable.
f. Life sciences and biomedical domain.
Data becomes a first-class, machine-accessible object as digital Evidence.
Major Outputs

a. Identifiers: harmonization CDL / EBI.
b. Publishers: roadmap to data citation.
c. Repositories: implement landing page metadata for data citation.
d. FAQs: guidance for common implementations based on JDDCP.
Some Participants


- European Bioinformatics Institute, National Library of Medicine, Dryad, FigShare, Dataverse.

- Harvard University, Columbia University, UCSD

- CrossRef, DataCite, California Digital Library
Participants

And you!
Identifier Harmonization Group

- California Digital Library (EZID / Name2Thing)
- European Bioinformatics Institute (identifiers.org)
- Co-representation from ELIXIR, BioCADDIE, NIH
- Harmonize identifier resolution for all standard bioinformatics databases across EU & US
- Workshop @ Harvard on June 2
DCIP Identifiers Workshop, June 2, 2016, Harvard University, Cambridge MA
John Kunze (CDL), Niall Beard (Manchester), Tim Clark (Harvard), Nick Juty (EBI), Ian Fore (NIH), Julie McMurry (UCSB), Jeff Grethe (UCSD), Rafa Jimenez (ELIXIR), Sarala Wimalaratne (EBI)
Prefix-Based Collection Access
draft-kunze-prefixes-00

Abstract

This document specifies a YAML [YAML] file that serves as an open registry of unique collection prefixes. These prefixes can be used by meta-resolvers to redirect identifiers to appropriate collection resolvers.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 9, 2016.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe the rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License.
Early Adopter Repositories

- Leads: Martin Fenner & Mercè Crosas

- Workshop **June 22 @ UCSD** precedes BioCADDIE Repositories Outreach meeting.

- Goal: develop proposed landing page metadata and outreach plan for repository adoption.

- Also Discuss - extension of metadata work to schema.org.
Publishers

- Leads: Amye Kenall & Helena Cousijn
- Outreach to other publishers in progress.
- Workshop **July 22 @ SpringerNature (London)** to help develop Publishers Roadmap for data citation.
FAQ/Outreach

- Leads: Joan Starr & Maryann Martone
- Building on the work from the other groups
- Materials to support the early adopters
Data Citations: A Primer

What Does a Data Citation Look Like?

Data citations look like other citations, but they reference data sets, not papers. But it’s perfectly fine to reference both a data set and the paper it came from, if appropriate. Here is how data might me cited in a (made-up) study

A (fake) study on lower back pain

To study the neural processing of lower back pain, we analyzed an fMRI data set from individuals exhibiting chronic lower back pain (Vrana et al., 2015a) from the study by Vrana et al. (2015b). And, just to put in a second example from a different repository, we utilized the replication dataset from Cranmer et al., (2016a) to test our network analysis.

References


DCIP Executive

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

- Goal: significantly improve biomedical translation.

- BioCADDIE / FORCE11 data citation pilot will promote implementing data citation in journals at scale.