Data Citation

From Principles to Implementation

DataCite

Martin Fenner

DataCite Technical Director

http://orcid.org/0000-0003-1419-2405

Joint Declaration of Data Citation Principles

https://www.force11.org/datacitation

DC^1

Data Citation Principles

Fun Fact

Joint Declaration doesn’t follow its own principles, e.g. for credit and attribution, and for persistent identifiers.

When citing please use: Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/datacitation].

Further Reading

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., et al. (2015). Achieving human and machine accessibility of cited data in scholarly publications. PeerJ. Computer Science, 1(9), e1. http://doi.org/10.7717/peerj-cs.1

1. Importance

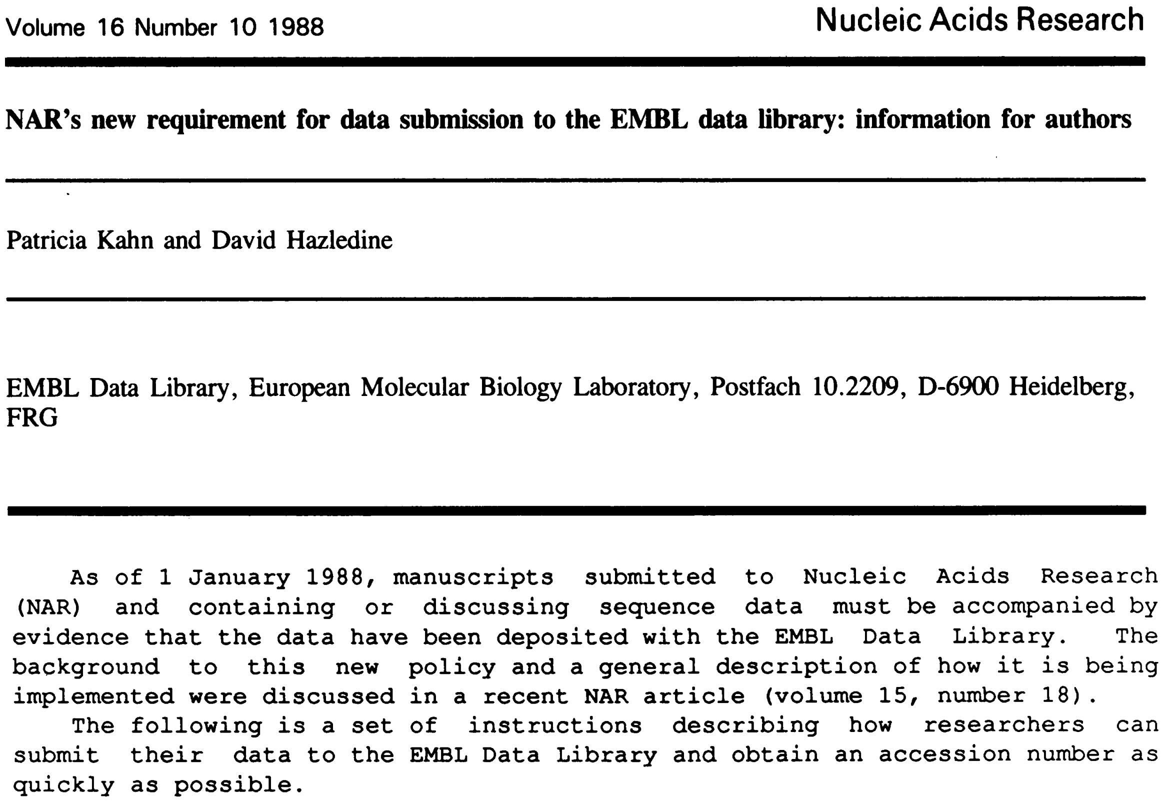
Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publication.

Data citation has a long tradition in the life sciences

My first data citation is from October 1996

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. U65091-U65093).

http://doi.org/10.1073/pnas.93.22.12298



Volume 16 Number 10 1988 Nucleic Acids Research

NAR's new requirement for data submission to the EMBL data library: information for authors

Patricia Kahn and David Hazledine

EMBL Data Library, European Molecular Biology Laboratory, Postfach 10.2209, D-6900 Heidelberg, FRG

As of 1 January 1988, manuscripts submitted to Nucleic Acids Research (NAR) and containing or discussing sequence data must be accompanied by evidence that the data have been deposited with the EMBL Data Library. The background to this new policy and a general description of how it is being implemented were discussed in a recent NAR article (volume 15, number 18).

The following is a set of instructions describing how researchers can submit their data to the EMBL Data Library and obtain an accession number as quickly as possible.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC336623/

The requirement to deposit nucleotide sequence data in public databases prior to manuscript submission has over time been extended to other data such as protein structures and gene expression data.

Extending this to all data underlying the conclusions of a paper is not (yet) an established community practice in the life sciences.

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.

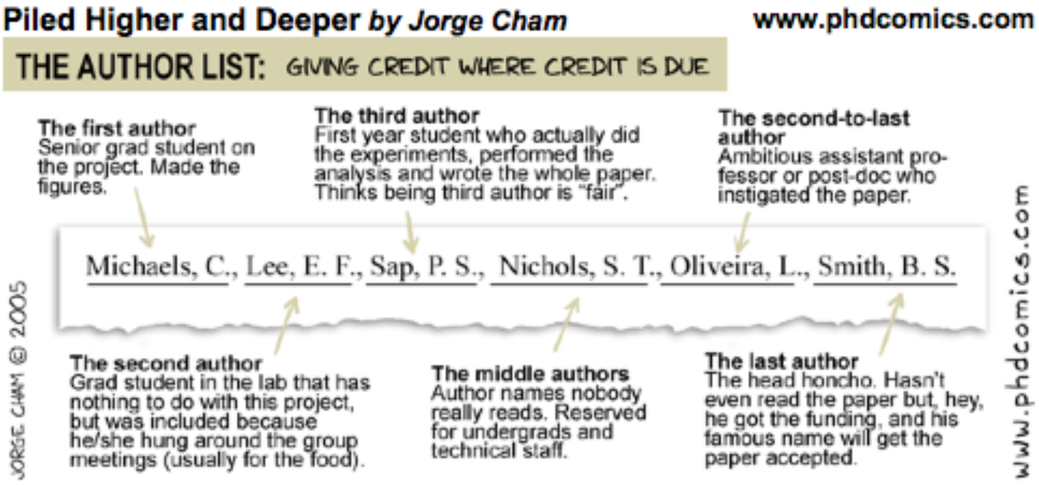
Fun Fact

Credit comes before evidence in Joint Declaration of Data Citation Principles.

Does CC0 require others who use my work to give me attribution?

No, and that's a big difference between CC0 and our licenses. Unlike our licenses, there are no conditions contained in CC0. Just like anything in the public domain, it will be possible for others to use or adapt it however they wish without attribution. However, this does not mean that you cannot request attribution in accordance with community or professional norms and standards.

https://wiki.creativecommons.org/wiki/CC0_FAQ



Piled Higher and Deeper *by Jorge Cham*

www.phdcomics.com

THE AUTHOR LIST: GIVING CREDIT WHERE CREDIT IS DUE

The first author
Senior grad student on the project. Made the figures.

The third author
First year student who actually did the experiments, performed the analysis and wrote the whole paper. Thinks being third author is "fair".

The second-to-last author
Ambitious assistant professor or post-doc who instigated the paper.

Michaels, C., Lee, E. F., Sap, P. S., Nichols, S. T., Oliveira, L., Smith, B. S.

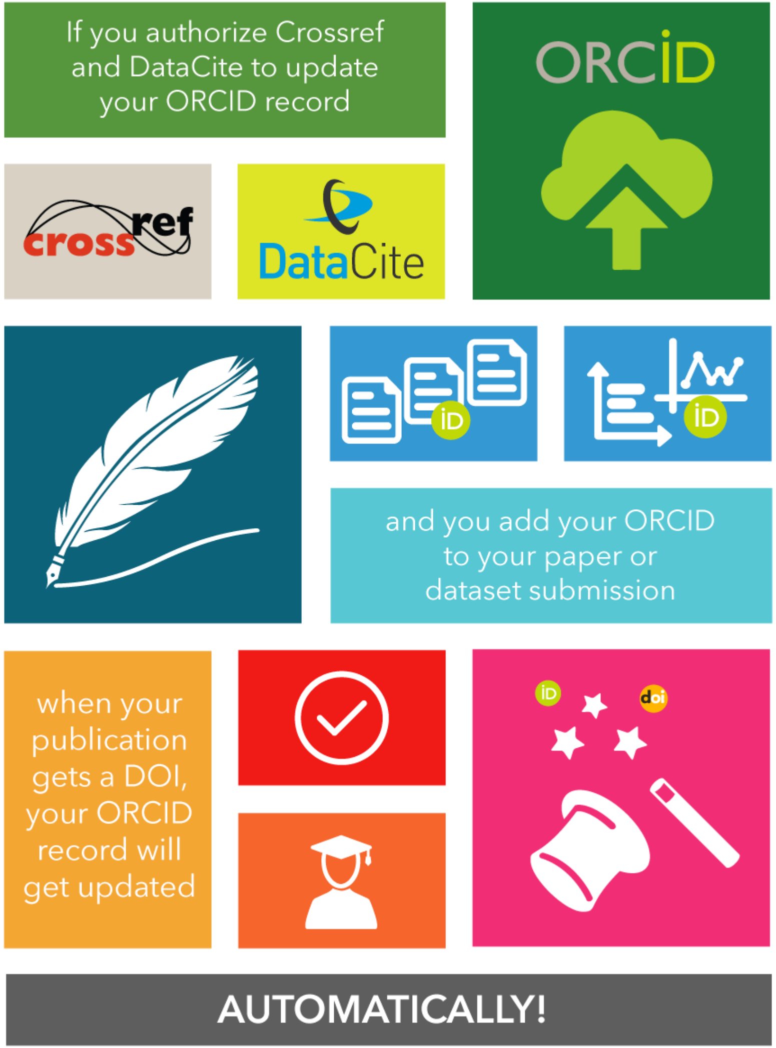
The second author
Grad student in the lab that has nothing to do with this project, but was included because he/she hung around the group meetings (usually for the food).

The middle authors
Author names nobody really reads. Reserved for undergrads and technical staff.

The last author
The head honcho. Hasn't even read the paper but, hey, he got the funding, and his famous name will get the paper accepted.

title: "Author List" - originally published 3/13/2005

http://www.phdcomics.com/comics/archive.php?comicid=562



If you authorize Crossref and DataCite to update your ORCID record

and you add your ORCID to your paper or dataset submission

when your publication gets a DOI, your ORCID record will get updated

AUTOMATICALLY!

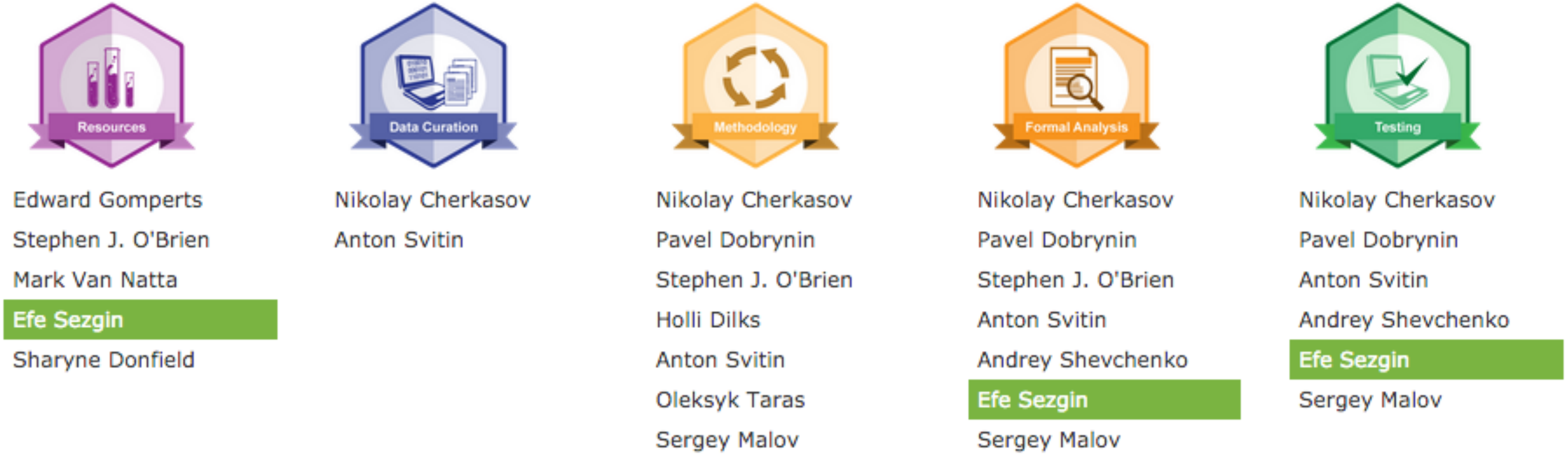
http://orcid.org/blog/2015/10/26/auto-update-has-arrived-orcid-records-move-next-level

Author Contributions

Conceived and designed the experiments: PD AMN. Performed the experiments: AMN BW AKH. Analyzed the data: AMN AKH JR MB PD. Contributed reagents/materials/analysis tools: JR MJ MB. Wrote the paper: AMN PD.

http://doi.org/10.1371/journal.pgen.1005087

Project CRediT



Resources	Data Curation	Methodology	Formal Analysis	Testing
Edward Gomperts	Nikolay Cherkasov	Nikolay Cherkasov	Nikolay Cherkasov	Nikolay Cherkasov
Stephen J. O'Brien	Anton Svitin	Pavel Dobrynin	Pavel Dobrynin	Pavel Dobrynin
Mark Van Natta		Stephen J. O'Brien	Stephen J. O'Brien	Anton Svitin
Efe Sezgin		Holli Dilks	Anton Svitin	Andrey Shevchenko
Sharyne Donfield		Anton Svitin	Andrey Shevchenko	Efe Sezgin
		Oleksyk Taras	Efe Sezgin	Sergey Malov
		Sergey Malov	Sergey Malov	

http://casrai.org/CRediT

http://www.gigasciencejournal.com/content/3/1/18/about#open-badges

3. Evidence

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

Joint Data Archiving Policy (JDAP)

[Journal] requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as [list of approved archives here]. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to a year after publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species.

http://datadryad.org/pages/jdap

Beyond the PDF

A text document with tables, figures and references is no longer the right format to adequately describe research.

We don’t have a common document format for this yet, but it should be a container format that can hold multiple file types, and has rich metadata, including strong citation support.

JATS could become (or could support) this common document format.

4. Unique Identification

A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

Figure 3. Multiple Alignment of Ten Conserved Motifs in the RAG1 Core Proteins and *Transib* TPases

The motifs are underlined and numbered from 1 to 10. Starting positions of the motifs immediately follow the corresponding protein names. Distances between the motifs are indicated in numbers of aa residues. Black circles denote conserved residues that form the RAG1/*Transib* catalytic DDE triad. The RAG1 proteins are as follows: RAG1_XL (GenBank GI no. 2501723, Xenopus laevis, frog), RAG1_HS (4557841, *Homo sapiens,*human), RAG1_GG (131826, *Gallus gallus,* chicken), RAG1_CL (1470117,Carcharhinus leucas, bull shark), RAG1_FR (4426834, Fugu rubripes, fugu fish).

not machine actionable, not globally unique

http://doi.org/10.1371/journal.pbio.0030181

Antibodies.

The antibodies used in this study included the following: rabbit polyclonal antibodies to GABA$_A$ receptor α2 (catalog #600-401-D45 RRID:AB_11182018; Rockland Immunochemicals), α5 (catalog #AB9678 RRID:AB_570435; Millipore), β3 (catalog #ab4046 RRID:AB_2109564; Abcam), γ2 (extracellular epitope, catalog #224 003 RRID:AB_2263066; Synaptic Systems), and AMPA receptor GluA1 (catalog #AB1504 RRID:AB_2113602; Millipore; and extracellular epitope, catalog #PC246-100UG RRID:AB_564636; Millipore) …

need context to be machine actionable, not globally unique

http.//doi.org/10.1523/JNEUROSCI.4415-13.2014

17. Yim KM, Ng HW, Chan CK, Yip G, Lau FL. Sibutramine-induced acute myocardial infarction in a young lady. Clin Toxicol (Phila). 2008; 46(9):877-879.
18. Waszkiewicz N, Zalewska-Szajda B, Szajda SD, Simonienko K, Zalewska A, Szulc A et al.. Sibutramine-induced mania as the first manifestation of bipolar disorder. BMC Psychiatry. 2012; 12:43.
19. Yet Another DataTables Column Filter. https://github.com/vedmack/yadcf

not persistent

http.//doi.org/10.1186/s13321-015-0077-3

Data access

The high-throughput read data is deposited at the European Nucleotide Archive (ENA) with the accession no. PRJEB7268 (http://www.ebi.ac.uk/ena/data/view/PRJEB7268).

good, but what URL?

(http://www.ncbi.nlm.nih.gov/bioproject/PRJEB7268/)

http://doi.org/10.1371/journal.pgen.1005087

Recommendation

- Use persistent identifier expressed as URI,
  e.g. http.//doi.org/10.1186/s13321-015-0077-3.
- Always include basic metadata, e.g. authors, title, publication date and publication venue.
- Put all citations into the reference list and make these metadata available in machine-readable format

The Importance of Reference Lists

- additional metadata beyond the unique identifier that can provide context
- facilitate extraction of machine readable metadata compared to embedding unique identifiers directly in article text
- access to article text with embedded unique identifiers might not be available if not open access

Updates to JATS to better support data citation

- two new elements: <version> and <data-title>
- new attribute @assigning-authority for elements <ext-link> and <pub-id>
- “data” as a suggested value for attribute @publication-type
- new value for attribute @person-group-type, for the data curator
- additional identifier values for the @pub-id-type attribute

Added in JATS 1.1d2

http://www.ncbi.nlm.nih.gov/books/NBK280240/

5. Access

Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

It is important that data are cited using machine-actionable unique identifiers, which in general means URIs.

These URIs should resolve to a landing page that holds human and machine-readable information about the resource.

Content negotiation and links in HTTP headers can be used to resolve the URI directly to the dataset in a machine-readable way.

http://doi.org/10.7717/peerj-cs.1

6. Persistence

Unique identifiers, and metadata describing the data, and its disposition, should persist – even beyond the lifespan of the data they describe.

Metadata for data that have been cited should persist.

Not all research data and their metadata can or should persist.

Metadata for most data that have been published should persist.

DOI names are persistent identifiers with focus on citation and publishing workflows.

Other identifiers might be more appropriate if data are not persistent, or used in a different context.

Data can have more than one identifier.

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.

For evidence we want to cite data as granular as possible.

For credit we want to cite data as broadly as possible.

FRBR - Functional Requirements for Bibliographic Records

Ontology to describe different representations of a work

- work
- expression
- manifestation
- item

http://archive.ifla.org/VII/s13/frbr/frbr1.htm

Challenges with Specificity

- versions
- slicing of fixed but large data
- dynamic data

For dynamic data the RDA Working Group on Data Citation recommends timestamped queries, but discussion is still ongoing

https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_150924.pdf

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.

We recognise that the challenges associated with data publication vary across disciplines, and we encourage research communities to develop citation systems that work well for them. Our recommended format for data citation is as follows:

Creator (PublicationYear): Title. Publisher. Identifier

It may also be desirable to include information about two optional properties, Version and ResourceType (as appropriate). If so, the recommended form is as follows:

Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

https://www.datacite.org/services/cite-your-data.html

Potential Implementation Differences

Citation Styles

Very few styles (e.g. APA) specifically support data citation. The NLM style recommendation for data citation is from 2007.

Separate reference lists

Some journals (e.g. Scientific Data) use separate reference lists for data. Not all references need to go into the PDF version of a publication.

Identifiers for collections

Rather than citing every single dataset in a publication (or listing them in the supplementary information), we can assign persistent identifiers with metadata to collections, and cite those.

http://www.ncbi.nlm.nih.gov/books/NBK7273/#A57573

More work needs to be done to bring data citation from principles to implementation.