Linking Articles to Data

J. Hourclé
joseph.a.hourcle@nasa.gov

W. Chang
wchang@nist.gov

F. Linares
falinares@iiaweb.com

G. Palanisamy
palanisamyg@ornl.gov

B. Wilson
wilsonbe@ornl.gov

There is an obvious need for scientific articles to cite the data being used, but should those citations link directly to the scientific data?

*We present recommendations for '**data landing pages**', an interstitial document with information about the scientific data being published by a science archive or cited by a paper, and their advantages over linking directly to scientific data in a data citation or using a formally published article as a proxy for the data citation. We propose two complementary forms, one provided and maintained by the data archive and a '**citation page**' that an author can attach to an article as supplementary material.*

The *Data Landing Page*

The group supplying or publishing the data should provide a page of information about the data. There should be standards created for these pages that have minimal information necessary for basic citation use, but that is extensible for domain-specific information or other value-added services, such as linking to papers that use the data or embedding tools to interact with the data.

Acts as the endpoint for citations

Citations should go to this intermediary ***landing page***, rather than directly to the data. Sending researchers directly to the data data can be a disservice, as the data may not be useful on its own without the proper software to read it or the proper documentation to understand it. The data may be excessive in size, and pushing researchers to download data without an interstitial warning about what they are downloading is a disservice to both interested researchers and to the providers hosting the data.

Contains information to generate citations

The ***landing page*** should have information about the data suitable to create a citation in the various different citation standards. The page should identify the common ***Dublin Core*** attributes, such as 'author' and 'title', to ensure the data is cited consistently across disciplines who may attribute 'authorship' to different roles in the data creation: the PIs, the software author, the pipeline manager, or even the instrument. Each discipline may have their own standards for data citation, but we also recommend using discipline agnostic standards, such as ***DataCite***.

For those disciplines where citation of data is not yet an accepted practice, the landing page can also include acknowledgement text. Both the acknowledgement and citation should include the DOI to this landing page to enable data centers to more easily find articles published using the data.

Can be updated

The ***landing page*** may change over time. Unlike using a journal article as a proxy citation for the data, we expect the landing page to be updated as the data are better understood. If the data are moved between archives, mirrored, or repackaged, the page should be updated to reflect this. Additional information may be added, such as recommendations for new tools that could be used to visualize the data that may not have existed when the data was first released.

Must be long-term

The ***landing page*** must be stored for the long-term, and must have a DOI assigned to it. Even if the data is no longer available, the landing page should remain to explain why.

If that specific version of the data is no longer available because it was deprecated by some other data, the page should link to a ***'versionless' landing page***–one describing the overall collection without discriminating on versions–that would then link to the most recent edition of the data. Although it may be appropriate to link to the next version of the data, we recommend using the 'versionless' page to prevent users from needing to follow dozens of links for frequently recalibrated data.

Should have provisions for machine readability

The ***landing page*** should provide a mechanism for machine parsing. The page may have links to the metadata encoded in various formats, such as by using an HTML "`<link rel='resourcemap'>`" element to point to an ***OAI-ORE Resource Map***. Archives may use ***HTTP Content Negotiation*** to serve either human or machine parsable files as appropriate. They could also always serve an XML representation and use XSLT to transform it for human-readability.

Each discipline may have different metadata requirements and standards, so it may be necessary to link to multiple metadata records or encode them into into a single XML file with multiple namespaces. We hope that over time, the science community will agree on discipline agnostic standards for provenance, basic identity and other non-discipline specific metadata to reduce the need to support multiple competing standards.

Should link to the data

If the data is still available, the ***landing page*** should have information on how to obtain the data. Where possible, these should be links to the data itself, but it may also be contact infomation to coordinate having the data sent on physical media for extremely large datasets. They may link to search systems for the dataset to allow subsetting, filtering, or other reduction of the data.

Should link to documentation

The ***landing page*** should provide references to any documentation necessary for understanding or using the data. This may be journal articles or grey literature describing the processing of the data, the design of the instrument that collected the data, or the software necessary to read, calibrate or otherwise use the data.

Should link to data on which it is based

If the data being described by the ***landing page*** is also based upon some published data, it should cite that data and provide a link to the appropriate landing page. As standards evolve, it may be possible to describe the full provenance chain.

May credit any number of people

As different citation formats are going to include or neglect different aspects of the data generation pipeline, the data landing page can give proper credit to any of the people involved in the data's collection, processing, validation, maintenance or other related tasks.

May be extended as desired

We only describe the minimum necessary for a ***data landing page***. Archives could extend them any number of ways, such as adding plots, images or movies to visualize the data. They could embed tools to allow scientists to directly interact with the data, such as rotating three dimensional structures or sorting through catalogs or other tabular data; this allows ***enhanced publications*** or ***data interactive publications*** in situations where the journal publisher doesn't directly support these features.

The *Data Citation Page*

As the ***data landing page*** provides us with the author, title and other necessary information, it is now possible to construct all components of the citation except for the subsetting. As each discipline may have different rules for subsetting, we propose that in the case of significant processing or subsetting, the author create a similar document that we call the ***data citation page***.

Acts as an extended methods supplement

The ***citation page*** provides information about the methods used to calibrate, correct, reduce, harmonize, manipulate or otherwise process the data. It may make references to specific software and tools used, or just clarify the nature of the processing that has been applied.

Must be long-term

The citation page should be archived for as long as the article itself. Ideally, the citation page would be archived as a supplement to the published article. If the publisher has no provisions for supplemental materials, it may be stored in a discipline or institutional repository, or other archive. As a last resort, it can be distributed through the author's website.

May aggregate data from multiple sources

For research that uses data from a large number of sources, the citation page can be used to list and link to all of the individual sources, rather than citing each data set individually.

Links to the data source

The ***citation page*** should provide the DOI of the ***landing page(s)*** for the data that was used.

May be domain specific

Although we hope that there will develop domain-agnostic standards for describing the provenance of data, we recognize that each discipline will process their data differently and use different language to describe it. We assume that the ***citation page*** is written for the same audience as the paper, and will be written in free-form text unless the domain has an accepted standard for describing data provenance.

The *Data Citation*

Once we have the information from the ***data landing page*** and optional ***data citation page***, we can use these to construct a ***data citation***.

Has more than one format

The ***data citation*** should be in the format proscribed by the style guide used by the journal the article is being published in.

Must include the Landing Page DOI

The DOI allows data centers to be able to find articles published using their data, as the citation format will change based on the style guide. We recommend using the HTTP URL form (`http://dx.doi.org/10...`) as many publishers don't turn the doi-prefixed form into a hyperlink.

If the author is using many data sources, the author should instead aggregate those DOIs within a citation page, and reference the ***citation page*** instead.

May include a link to a Data Citation Page

If significant subsetting or processing has been applied to the data, the author should create a ***data citation page*** to describe what transformations have been applied to the data, and reference it from the citation.

Acknowledgements

The recommendation for linking to landing pages came out of the Technical Breakout from the Board of Research Data & Information's August 2011 meeting on 'Developing Data Attribution and Citation Practices and Standards'. We would like to thank the other members of that discussion: Paul Groth, Allen Renear, Herbert van de Sompel, and Martie Van Deventer. We also thank the other participants in the BRDI meeting who gave feedback to this proposal.

References

Ball, A. & Duke, M., (2011). "Cite Datasets and Link to Publications". `http://www.dcc.ac.uk/resources/how-guides`

DataCite, (2011). "DataCite Metadata Schema for the Publication and Citation of Research Data". `http://dx.doi.org/10.5438/0005`

Domenico, (2011). "White Paper on Data Interactive Publications". `https://sites.google.com/site/datainteractivepublications/home/white-paper-on-data-interactive-publications`

OAI, (2008). "ORE User Guide - HTTP Implementation". `http://www.openarchives.org/ore/1.0/http.html`

Parsons M. & ESIP, "How to Cite an Earth Science Data Set"

SURF Foundation, "Enhanced Publications" `http://surffoundation.nl/enhancedpublications`

Handout and poster available at:

`http://docs.virtualsolar.org/wiki/Citation`