Human & Machine Actionable Data Citations

J.A. Hourclé
NASA-GSFC (Wyle)
joseph.a.hourcle@nasa.gov

R.R. Downs
CIESIN, Columbia University
rdowns@ciesin.columbia.edu

R. Duerr
NSIDC & Data Conservancy
rduerr@nsidc.org

DC[1]

[1]Data Citation Principles

Last year, the Joint Declaration for Data Citation Principles was released. [1] The preamble of the declaration states "These principles recognize the dual necessity of creating citation practices that are both ***human understandable and machine-actionable***."

A group was formed to discuss the issues regarding implementing these principles and currently has a paper in process, "***Achieving human and machine accessibility of cited data in scholarly publications***"[2] with a proposal of how to achieve the goal in the preamble.

We present a short overview of both the data citation principles and our proposal. ***We invite comments on the paper during its review period.***

References:

[1] 2014, "Joint Declaration of Data Citation Principles", http://www.force11.org/datacitation

[2] (in process), "Achieving human and machine accessibility of cited data in scholarly publications", https://peerj.com/preprints/697/

[3] 2012, "Data Citation Guidelines for Data Providers and Archives", http://commons.esipfed.org/node/308

[4] 2013, "DOIs Should Not Link to Data", http://dx.doi.org/10.5281/zenodo.10057

[5] 2014, "Data Catalog Vocabulary (DCAT): W3C Recommendation", http://www.w3.org/TR/vocab-dcat/

[6] "What is ORCID?", http://orcid.org/node/47

[7] 1998, "RFC2295 - Transparent Content Negotiation in HTTP", https://www.ietf.org/rfc/rfc2295.txt

[8] 2010, "RFC5988 - Web Linking", https://www.ietf.org/rfc/rfc5988.txt

[9] 2008, "ORE User Guide - Resource Map Discovery", http://www.openarchives.org/ore/1.0/discovery

[10] 2012, "For Attribution–Developing Data Attribution and Citation Practices and Standards", http://www.nap.edu/catalog.php?record_id=13564

Acknowledgements:

The methods for machine actionability derive from discussions of the Technical Breakout at the BRDI meeting, 'For Attribution - Developing Data Attribution and Citation Practices and Standards'. [10]

We would like to thank FORCE11 for hosting and organizing the tasks groups that have generated the referenced documents [1,2]. We would also thank the many people who have volunteered their time and expertise on the Data Citation Principles Synthesis Group and the Data Citation Implementation Group.

Joint Declaration of Data Citation Principles:

See the accompanying handout or the FORCE11 website for the official text of the principles. Thus far, the recommendations are compatable with the ESIP Data Citation Guidelines. [3]

We have paraphrased the principles to explain some wording subtleties and draw out ***issues of importance*** for our community. ***Highlighted items*** are specific details from the paper.

Background:

Reprocibility of science ***relies on knowing the evidence*** used.

Producers of data should be ***given credit*** for their contributions.

We need ***cross-discipline standards*** for citing data, as data used by communities other than the one that collected it.

Implementations will ***vary by discipline*** and ***evolve over time***.

Principle 1: Importance

The data used as evidence should be given at least as much credit as articles used to set the context for the paper.

Communities should consider a person's work in producing good data for others to use when considering tenure, promotion & grants.

Principle 2 : Credit and attribution

There is no simple 'author' for data, and citing a 'first results' or 'instrument' paper doesn't give proper credit to people who may come in later and give significant contributions to the calibration or other understanding of the data.

Principle 3 : Evidence

The ***data used to support your research should be cited in the reference list***. You should also ***link the data being used as evidence near the claim being made***; depending on the journal, this may be inline text, a footnote, or a caption to a plot or table.

Principle 4 : Unique Identification

For this whole system to work, we need ***cross-discipline identifiers***.

Although many would like to standardize on DOIs (Digital Object Identifiers), ***many groups have standardized on other identifier schemes and so the paper recommends "any currently-available identifier scheme that is machine actionable, globally unique, and widely (and currently) used by a community; and that has a long term commitment to persistence".***[2] DOIs would allow us to use existing bibliographic tools to track the use of data, ***reduce the work needed to prepare for Senior Reviews***, and ***find uses of our data by other communities***.

To be 'machine actionable', ***identifiers should be a fully qualified URL*** to a resolver for the identifier. (i.e., 'http://dx.doi.org/...', not 'doi:...').

Principle 5 : Access

Citations do not need to (and should not, in our community[4]) link directly to the data. ***Identifiers should point to a landing page or set of pages rather than to the data itself*** so that people can make an informed choice before potentially downloading terabytes of data that isn't useful to them.

These 'landing pages' can be updated to ***provide links to current documentation, software, related data*** (eg. alternate processed forms or from complementary missions), published papers using the data, and whatever metadata the community feels is appropriate for that data.

Principle 6 : Persistence

Even if the data goes away (replaced by better data, removed due to security or budget, or lost by accident), the 'landing page' remains, so we ***do not have a gap in the scientific record***.

When appropriate, this 'tombstone page' should describe why data was removed, and link to possible replacements or alternatives (eg, better calibrated versions).

Principle 7 : Versioning and granularity

If there are formal releases, assign an identifier to each one, so researchers can ***cite a specific version***. If under continuous release, citations should ***include an access date***.

If you didn't analyze all of the data, ***describe what portion you used*** (i.e., date, spectral or spatial ranges; specific observing modes; or any other filtering or subsetting.)

RDA is working on interop. subset standards; Thurs@1:30pm: Workshop on 'Dynamic Data Citation'

Principle 8 : Interoperability and flexibility

Every journal / community cites things a little bit differently, and has different metadata requirements. The data citation community is working towards a ***universal framework*** that ***the each community can extend for their specific needs***.

Landing Pages:

The identifier included in a citation should ***point to a landing page or set of pages rather than to the data itself***.

Although the paper's stated goals for this are to deal with access restrictions, persistence, and to provide for multiple packaged forms of the data, this indirection is necessary when dealing with large earth science datasets. This allows us to cite collections of any size, that may not be online (eg, physical samples, embargoed, or in dark archives), or no longer exist (older calibrations).

These pages should have whatever information is appropriate for their community, but at a minumum, they should have the appropriate metadata to enable someone using the data to create a citation string:

- ***Dataset Identifier***
- ***Title***
- ***Creator***
- ***Publisher or Contact***
- ***Release Date or Year***
- ***Version***

They must also contain a description of the dataset; use ***W3C's DCAT vocabulary*** [5] for interoperability, but also use the appropriate metadata standards for the data's intended communities.

Landing pages should also ***include the license under which the data is released, a persistence statement, and identifiers, such as ORCID*** [6], to provide attribution to the individuals and organizations that contributed to, curated, and maintain the data. These pages may provide additional information about the data such as context, caveats, links to software and documentation, information on data availability, or whatever the publisher feels adds value to the data.

Human & Machine Actionable:

As the identifiers are URLs, we can use existing web techniques to provide both HTML for humans and an alternate format for machines:

1. **HTTP Accept Headers:**

   Use content negotiation to serve HTML or alternative machine-readable format(s). [7] This allows user agents to specify that format they would prefer, and the web server to automatically serve the requested format.

2. **HTTP Link Headers**

   Use web linking to have the web server to inform the user-agent what alternate formats are available. [8,9]

   `Link: uri-to-an-alternate; rel="alternate"; media="application/xml"`

3. **HTML Link Elements**

   Use HTML elements to embed web linking information:

   `<link href="uri-to-an-alternate" rel="alternate" type="application/xml">`

We recommend a combination of the three; all have advantages, but none are clearly superior on their own:

	#1	#2	#3
Discovery of alternates	*No*	**Yes**	**Yes**
Supported for all file types	**Yes**	**Yes**	*No*
Remains attached if downloaded	*No*	*No*	**Yes**
Doesn't require multiple requests	**Yes**	*No*	*No*
Can support multiple languages	**Yes**	**Yes**	**Yes**
Can support multiple of the same mime-type	~	**Yes**	**Yes**

As content negotiation [7] operates primarily on MIME type, it is more difficult to differentiate between two different schemas that are both serialized as XML without coining new types (e.g., 'application/atom+xml' vs. 'application/xhtml+xml'). Web linking [8] allows for alternate relationship types which could be used to link to specific schema. Another solution would be to use web linking to link to an ORE resource map [9] that could more explicitly describe the relationships between alternatives.

This poster can be downloaded or commented on at:
http://commons.esipfed.org/node/7768

To download or comment on the paper :
https://peerj.com/preprints/697/