Data Citation Then and Now

Mark A. Parsons with help from Ruth Duerr and Peter Fox

17 June 2014
GeoData 2014
Boulder, CO
The Evolution of Data Citation

- Data was part of the literature—tables, maps, monographs, etc.—and we cited accordingly. (Some data were still hoarded).
- Digital data becomes the norm. It’s messier and we forget how to do cite it routinely.
- Initial efforts to define digital data citation in the 90s - early 00s
  - Right idea, little traction
  - Partially conflated with the citing URLs issue
- A blossoming in the mid-late 00s.
  - Multiple disciplines start developing approaches and guidelines
  - DOI a big driver, especially for DataCite, but other identifiers used too (Handles, LSIDs, UNFs, ARKs and good ol’ URI/Ls)
  - A slightly competitive atmosphere
- GeoData 2011 a milestone for **ESIP Citation Guidelines**—most sophisticated to date. ([http://dx.doi.org/10.7269/P34F1NNJ](http://dx.doi.org/10.7269/P34F1NNJ))
The Evolution of Data Citation

• **Now a consensus phase**
  
  - *Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data*. 2013.  
    [http://dx.doi.org/10.2481/dsj.OSOM13-043](http://dx.doi.org/10.2481/dsj.OSOM13-043)
  
  - Draft Global Joint Declaration of Data Citation Principles. 2013.  
The Evolution of Data Citation—Next

- **Implementation phase just begun**
  - ESIP Guidelines adopted by a variety of NASA and NOAA data centers and internationally by GEOSS.
  - AGU Publishing Committee is developing author guidelines based on ESIP.
  - Other disciplines, notably social science, has relationships with publishers.
  - Several data centers partnering with publishers, e.g. Elsevier’s “article of the future”.
  - New PLOS data policy.
  - Joint engagement activity following on the joint principles.
- It happens locally and requires culture change so debates will continue.
Outline of 2011 Talk

• Purpose of Data Citation
• How it’s currently done
• Basic citation form and content
• Identifiers and locators
• Microcitation
Purpose of data citation
Purpose of Data Citation

• Credit for data creators and stewards
• Track impact of data set
• Accountability for creators and stewards
• Aid reproducibility through direct, unambiguous connection to the precise data used

• A location/reference mechanism not a discovery mechanism per se.
Purpose of Data Citation

- Aid scientific reproducibility through direct, unambiguous connection to the precise data used.
- Credit for data authors and stewards
- Accountability for creators and stewards
- Track impact of data set
- Help identify data use (e.g., trackbacks)
  - Data authors can verify how their data are being used.
  - Users can better understand the application of the data.
- A locator/reference mechanism not a discovery mechanism per se
The Noble Eight-Fold Path to Citing Data

1. Importance
2. Credit and attribution
3. Evidence
4. Unique Identification
5. Access
6. Persistence
7. Specificity and verifiability
8. Interoperability and flexibility

Principles are supplemented with a glossary, references and examples
http://force11.org/datacitation
Purpose of Data Citation

A locator/reference/linking mechanism!

This helps

- Aid scientific reproducibility through direct, unambiguous connection to the precise data used
- Identify and track data use

It contributes but is NOT central to

- Credit for data authors and stewards
- Accountability for creators and stewards
- Tracking impact of data set
Is data publication the right metaphor?

M. A. Parsons & P. A. Fox

*Data Science Journal, 2013*

http://dx.doi.org/10.2481/dsj.WDS-042

Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature.

—Lakoff and Johnson, 1980

If we hear the same language over and over, we will think more and more in terms of the frames and metaphors activated by that language.

—Lakoff, 2008
How it’s currently done
How data citation is currently done

• Citation of traditional publication that actually contains the data, e.g. a parameterization value.
• Not mentioned, just used, e.g., in tables or figures
• Reference to name or source of data in text
• URL in text (with variable degrees of specificity)
• Citation of related paper (e.g. CRU Temp. records recommend citing two old journal articles which do not contain the actual data or full description of methods)
• Citation of actual data set typically using recommended citation given by data center
• Citation of data set including a persistent identifier/locator, typically a DOI
How data citation is currently done

- Citation of traditional publication that actually contains the data, e.g. a parameterization value.
- Not mentioned, just used, e.g., in tables or figures
- Reference to name or source of data in text
- URL in text (with variable degrees of specificity)
- Citation of related paper (e.g. CRU Temp. records recommend citing two old journal articles which do not contain the actual data or full description of methods.)
- Citation of actual data set typically using recommended citation given by data center
- Citation of data set including a persistent identifier/locator, typically a DOI
Then

- 2009: 1.7%
- 2008: 1.3%
- 2007: 0.9%
- 2006: 1.3%
- 2005: 0.7%
- 2004: 0.7%
- 2003: 1.0%
- 2002: 1.3%

Formal Citation
Total Entries

“MODIS Snow Cover Data” in Google Scholar
How data citation is currently done

- **Implementation phase just begun**
  - ESIP Guidelines adopted by a variety of NASA and NOAA data centers and internationally by GEOSS.
  - AGU Publishing Committee is developing author guidelines based on ESIP.
  - Other disciplines, notably social science, has relationships with publishers.
  - Several data centers partnering with publishers, e.g. Elsevier’s “article of the future”.
  - New PLOS data policy.
  - Joint engagement activity following on the joint principles.
- It happens locally and requires culture change so debates will continue.
Basic content
Basic data citation form and content

Per DataCite:
Creator. PublicationYear. Title. [Version]. Publisher. [ResourceType]. Identifier.

Per ESIP:
Author(s). ReleaseDate. Title, [version]. [editor(s)]. Archive and/or Distributor. Locator. [date/time accessed]. [subset used].
An Example Citation

Identifiers and locators
An assessment of identification schemes for digital Earth science data

<table>
<thead>
<tr>
<th>ID Scheme</th>
<th>Unique Identifier</th>
<th>Unique Locator</th>
<th>Citable Locator</th>
<th>Scientifically Unique ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL/N/I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PURL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XRI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UUID</td>
<td><strong>Good</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

An assessment of identification schemes for digital Earth science data

<table>
<thead>
<tr>
<th>ID Scheme</th>
<th>Unique Identifier</th>
<th>Unique Locator</th>
<th>Citable Locator</th>
<th>Scientifically Unique ID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data Set</td>
<td>Item</td>
<td>Data Set</td>
<td>Item</td>
</tr>
<tr>
<td>URL/N/I</td>
<td>Yellow</td>
<td>Red</td>
<td>Green</td>
<td>Yellow</td>
</tr>
<tr>
<td>PURL</td>
<td>Yellow</td>
<td>Red</td>
<td>Green</td>
<td>Yellow</td>
</tr>
<tr>
<td>XRI</td>
<td>Yellow</td>
<td>Red</td>
<td>Green</td>
<td>Yellow</td>
</tr>
<tr>
<td>Handle</td>
<td>Yellow</td>
<td>Red</td>
<td>Green</td>
<td>Yellow</td>
</tr>
<tr>
<td>DOI</td>
<td>Yellow</td>
<td>Red</td>
<td>Green</td>
<td>Yellow</td>
</tr>
<tr>
<td>ARK</td>
<td>Yellow</td>
<td>Red</td>
<td>Green</td>
<td>Yellow</td>
</tr>
<tr>
<td>LSID</td>
<td>Yellow</td>
<td>Red</td>
<td>Green</td>
<td>Yellow</td>
</tr>
<tr>
<td>OID</td>
<td>Yellow</td>
<td>Red</td>
<td>Green</td>
<td>Yellow</td>
</tr>
<tr>
<td>UUID</td>
<td>Yellow</td>
<td>Red</td>
<td>Green</td>
<td>Yellow</td>
</tr>
</tbody>
</table>


http://dx.doi.org/10.1007/s12145-011-0083-6
What needs an identifier/locator?
What needs to be cited?

- Everything needs an identifier. Most things need locators. Intellectual content needs citation.
- Different versions of things may need different identifiers/locators.
- Subsets may need identifiers or clear reference to sub-setting process (e.g. space and time).
- Different representations (conceptual models) may need different identifiers/locators. E.g. Maps.
Why the DOI?

- Not perfect but well understood by publishers
- Thomson Reuters collaborating with DataCite to get data citations in their index.

But...
- What is the citable unit?
- How do we handle different versions?
- What about “retired” data?
- When is a DOI assigned?
Versioning approach recommended by DCC

• “As DOIs are used to cite data as evidence, the dataset to which a DOI points should also remain unchanged, with any new version receiving a new DOI.”

• “There are two possible approaches the data repository can take: time slices and snapshots.”
When to assign a DOI?

• **First principle:** Data should be citable as soon as they are available for use by anyone other than the original authors.

• But...
  
  • Most people (falsely) believe that a DOI implies permanence so how do we cite transient data?
  
  • Some believe that a DOI should not be assigned until the data has undergone some level of review (e.g. Lawrence et al. 2010). So how do we cite data used before the review?
  
  • Data are often used by friends and collaborators in a raw, “unpublished” state. Should this use be cited with a DOI?
  
  • Near real time or preliminary data may only be available for a short uncurated period. There may not be a good match between the submission package and the distribution package. What gets the DOI? When?
Versioning and locators: some suggestions from NSIDC

- major version.minor version.[archive version]
- Individual stewards need to determine which are major vs. minor versions and describe the nature and file/record range of every version.
- Assign DOIs to major versions.
- Old DOIs should be maintained and point to some appropriate page that explains what happened to the old data if they were not archived.
- A new major version leads to the creation of a new collection-level metadata record that is distributed to appropriate registries. The older metadata record should remain with a pointer to the new version and with explanation of the status of the older version data.
- Major and minor version (after the first version) should be exposed with the data set title and recommended citation.
- Minor versions should be explained in documentation, ideally in file-level metadata.
- Applying UUIDs or ARKs to individual files upon ingest aids in tracking minor versions and historical citations.
Microcitation
Basic data citation form and content

Author(s). ReleaseDate. Title, Version. [editor(s)]. Archive and/or Distributor. Locator. [date/time accessed]. [subset used].

The best solution is to have unique identifiers or query IDs for subsets, but that won’t be available for most data sets for a long time, so we need alternative solutions...
Neither solution is perfect—‘locations’ or page numbers—because the problem is unsolvable. The best we can hope for is a choice...

Amazon’s Kindle will have page numbers that correspond to real books and locations by passage.
Chapter and Verse

- Bible
- Koran
- Bhagavad-Gita and Ramayana
- other sacred texts

- A “structural index”


“Just-in-time” citation

Approach being developed by an RDA Working Group

• Ensure data is time-stamped and versioned
• Assign PID to time-stamped query/selection expression
Content negotiation—the details of identifier resolution
So, you have a DOI, or a handle, ...

- Resolution service or URI?
- Landing pages (for metadata, citation recommendations, ...)
- Content-negotiation (conneg) for embedding in
  - Other pages
  - Applications
  - Ingest, object type repositories, communities
CLPX-Ground: ISA Snow Depth Transects and Related Measurements, Version 2

Documentation  FTP

This data set consists of snow depth data from nine study areas, within three larger-scale areas in northern Colorado (Fraser, North Park, and Rabbit Ears Meso-cell Study Areas (MSAs)). The study areas range from low-relief (flat topography) unforested areas with shallow snow covers, to high-relief (complex topography) densely forested areas with deep snow covers. Parameters measured include snow depth, snow wetness, snow temperature, soil temperature, and canopy cover. Photos were taken of surface roughness and terrain. Data were collected during February and March 2002 and 2003. This data set is part of the NASA Cold Land Processes Field Experiment (CLPX).

View Metadata Record

Data Citation
The following example shows how to cite the use of this data set in a publication. For more information, see our Use and Copyright Web page.


See Also
- CLPX Site
- Register as a CLPX user
- Contact User Services

“Landing Pages” For humans and machines
Landing page – a short form

http://data.rpi.edu/repository/handle/10833/24

KML files for Rensselaer Polytechnic Institute campus building footprints

Title: KML files for Rensselaer Polytechnic Institute campus building footprints

Author: Gill, William

Description: KML files (compatible with Google Earth and Google Maps) for footprints of 84 buildings on the Rensselaer Polytechnic Institute campus, created in August 2010. Data points were collected digitally using Google Maps aerial photography. Files downloaded on 2012-05-15 from http://www.rpi.edu/dept/cct/data/resources/buildings/

URI: http://hdl.handle.net/10833/24

Date: 2012-05-16

Files in this item

<table>
<thead>
<tr>
<th>Files</th>
<th>Size</th>
<th>Format</th>
<th>View</th>
</tr>
</thead>
<tbody>
<tr>
<td>_all_files.zip</td>
<td>45.09Kb</td>
<td>application/zip</td>
<td>View/Open</td>
</tr>
<tr>
<td>11th Street Boiler House.kml</td>
<td>935bytes</td>
<td>application/vnd.google-earth.kml+xml</td>
<td>View/Open</td>
</tr>
<tr>
<td>133 Sunset Terrace.kml</td>
<td>1.098Kb</td>
<td>application/vnd.google-earth.kml+xml</td>
<td>View/Open</td>
</tr>
</tbody>
</table>
KML files for Rensselaer Polytechnic Institute campus building footprints

<table>
<thead>
<tr>
<th>dc.contributor</th>
<th>Rensselaer Polytechnic Institute. DotCIO. Communications &amp; Middleware Technologies.</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc.creator</td>
<td>Gill, William</td>
</tr>
<tr>
<td>dc.date.issued</td>
<td>2012-05-16</td>
</tr>
<tr>
<td>dc.identifier.uri</td>
<td><a href="http://hdl.handle.net/10833/24">http://hdl.handle.net/10833/24</a></td>
</tr>
<tr>
<td>dc.description</td>
<td>KML files (compatible with Google Earth and Google Maps) for footprints of 84 buildings on the Rensselaer Polytechnic Institute campus, created in August 2010. Data points were collected digitally using Google Maps aerial photography. Files downloaded on 2012-05-15 from <a href="http://www.rpi.edu/dept/cct/data/resources/buildings/">http://www.rpi.edu/dept/cct/data/resources/buildings/</a></td>
</tr>
<tr>
<td>dc.format.extent</td>
<td>104 KB</td>
</tr>
<tr>
<td>dc.format.mimetype</td>
<td>application/vnd.google-earth.kml+xml</td>
</tr>
<tr>
<td>dc.rights</td>
<td>To reuse, modify, or share this data, contact Communications &amp; Middleware Technologies, Rensselaer Polytechnic Institute.</td>
</tr>
<tr>
<td>dc.subject</td>
<td>Rensselaer Polytechnic Institute campus building footprint data</td>
</tr>
<tr>
<td>dc.title</td>
<td>KML files for Rensselaer Polytechnic Institute campus building footprints</td>
</tr>
<tr>
<td>dc.type</td>
<td>Dataset</td>
</tr>
</tbody>
</table>
Content Negotiation—Conneg

• Many examples, but what follows is ~ from: http://www.crosscite.org/cn/

• What is it?
  – Es ce que vous parlez Français?
  – Do you speak html or JSON or RDF?

• For embedding in...
  – Other pages
  – Applications
  – Ingest, object type repositories, communities
Conneg

GET "Accept: text/html"
http://dx.doi.org/10.1126/science.169.3946.635

|   |
|   |
|   V

Publisher landing page
http://www.sciencemag.org/content/169/3946/635

GET "Accept: application/rdf+xml"
http://dx.doi.org/10.1126/science.169.3946.635

|   |
|   |
|   V

CrossRef metadata service
http://data.crossref.org/10.1126/science.169.3946.635
Supported content types..

<table>
<thead>
<tr>
<th>Format</th>
<th>Content Type</th>
<th>CrossRef</th>
<th>DataCite</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDF XML</td>
<td>application/rdf+xml</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RDF Turtle</td>
<td>text/turtle</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Citeproc JSON</td>
<td>application/vnd.citationstyles.csl+json</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Formatted text citation</td>
<td>text/x-bibliography</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RIS</td>
<td>application/x-research-info-systems</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>BibTeX</td>
<td>application/x-bibtex</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CrossRef Unixref XML</td>
<td>application/vnd.crossref.unixref+xml</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>DataCite XML</td>
<td>application/vnd.datacite.datacite+xml</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Further integration..
Get involved!

- RDA Working Group on citing dynamic data.
  - http://rd-alliance.org/working-groups/data-citation-wg.html
- RDA WG on identifier “types” and PID Interest Group
  - https://rd-alliance.org/working-groups/pid-information-types-wg.html
  - https://rd-alliance.org/internal-groups/pid-interest-group.html
- ESIP Preservation and Stewardship Committee
  - http://wiki.esipfed.org/index.php/Preservation_and_Stewardship
- Implementation Team for the Joint Citation Principles
  - http://www.force11.org/node/4849
Update of 2011 Talk

- **Purpose of Data Citation**—evolving into more specific concerns (slowly)
- **How it’s currently done**—consensus on needs and approach, but no substantial progress on implementation
- **Basic citation form and content**—basics are solid and generally consistent
- **Identifiers and locators**—consensus emerging, now looking at identifiers for everything and what to resolve.
- **Micro-citation**—good technical progress, hindered by social implementation and concept of citation vs. linking.
Overall Summary

• Use many metaphors and be cautious of them all.
• We know how to cite data (for the most part), we just need to make it a cultural practice. Just do it.
• Location and identity are different but can be the same. Separate concerns.
• Data citation is not literature citation. Micro-citation is more important and the citation needs to be machine interpretable and operable.
• Machine interpretable and operable.
• Everything in data stewardship needs an identifier, but… due diligence is the underlying requirement.
• These sort of socio-technical problems require collaborative, networked solutions. Participate.