Identity, Location, and Citation

Mark A. Parsons

USGS Digital Object Identifiers Focus Group
12 June 2013
First let’s talk about metaphor
Metaphor is for most people a device of the poetic imagination and the rhetorical flourish—a matter of extraordinary rather than ordinary language. Moreover, metaphor is typically viewed as characteristic of language alone, a matter of words rather than thought or action. For this reason, most people think they can get along perfectly well without metaphor.

We have found, on the contrary, that metaphor is pervasive in everyday life, not just in language but in thought and action. Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature.
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
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
Identifier vs. Locator

- **Human ID:** Mark Alan Parsons (son of Robert A. and Ann M., etc.)
  - every term defined independently (only unique in context/provenance)
  - Alternative like a social security number (or ORCID) requires a very well controlled central authority and unchanging objects.

- **Human Locator:** Amos Eaton Hall, Room 209, 110 8th St., Troy, NY 12180.
  - every term has a naming authority, i.e. a type of registry

- **Data Set IDs:** data set title, filename, database key, object id code (e.g. UUID), etc.

- **Data set Locators:** URL, directory structure, catalog number, registered locator (e.g. DOI), etc.
One of the main purposes of assigning DOI names (or any persistent identifier) is to separate the location information from any other metadata about a resource. Changeable location information is not considered part of the resource description. Once a resource has been registered with a persistent identifier, the only location information relevant for this resource from now on is that identifier, e.g., http://dx.doi.org/10.xx.

— DataCite Metadata Scheme for the Publication and Citation of Research Data, Version 2.2, July 2011 (my emphasis).
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
“MODIS Snow Cover Data” in Google Scholar
Data Citation Guidelines

  • Best available for Earth system science. Not yet widely adopted.
  • Best overall guide. Not yet widely adopted.
  • Detailed treatment of map-based data, but seemingly not well recognized. Does not address location.
• DataCite—a well-recognized consortium of libraries and related organizations working to define a citation approach and assign DOIs. Also working to get data citations included in citation indices.
• CODATA/ICSTI and NAS Task Group conducted an excellent workshop that summarizes approaches and issues: http://www.nap.edu/catalog.php?record_id=13564. A final report summarizing the state of the art will be out soon.
• DataVerse Network Project—a standard from the social science community using a Handle locator and “Universal Numerical Fingerprint” as a unique identifier.
• NASA DAACs, some NOAA and NSF centers adopting ESIP-based approaches.
• Various life and social science centers have standardized approaches with increasing adoption, e.g. Dryad.
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.
News Flash! September 2011
Greenland ice shrinks 15% since 1999 according to new edition of The Times Atlas
Headlines a couple days later

- Mapmakers' claim on shape of Greenland suddenly melts away
- A greener Greenland? Times Atlas 'error' overstates global warming
- Row over how much Greenland has shrunk
- Times Atlas is 'wrong on Greenland climate change'
- Times Atlas accused of 'absurd' climate change ice error
The Culprit?


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].

Richards, 2000. Does a detailed treatment for map-based data addressing issues of version and representation but does not explicitly address location.
Doing it as best we can...?


Get involved

• Research Data Alliance proposed Working Group on citing dynamic data.
  • http://rd-alliance.org/working-groups/data-citation-wg.html

• ESIP Preservation and Stewardship Committee
  • http://wiki.esipfed.org/index.php/Preservation_and_Stewardship
Data Set Identifiers: Applications and Implementation

R. Duerr
Outline

- ESIP activities
  - An assessment of identification schemes…
  - Best practices
  - Next steps
- DataCite vs CrossRef
An Assessment of identification schemes...

<table>
<thead>
<tr>
<th>ID Scheme</th>
<th>Data Set</th>
<th>Item</th>
<th>Data Set</th>
<th>Item</th>
<th>Data Set</th>
<th>Item</th>
<th>Data Set</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL/N/I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PURL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XRI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UUID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Presented to the USGS Digital Object Identifiers Focus Group
By Ruth Duerr, June 12, 2013
Best Practices

- DOI’s for data sets (for now)
- UUID’s for data items
- Systems need to be prepared to support multiple identifiers and locators over time
- ESIP data citation guidelines (http://commons.esipfed.org/node/308)
- Assign identifiers/locators to associated provenance and contextual materials
Next Steps

• What other types of materials need identification and what kinds of identifiers should they have?
  • People, organizations, etc.
  • Likely in conjunction with the RDA group on this topic
• Provenance and context content standard
• PROV-ES ontology profile for the Earth Sciences
DataCite/EZID

- Primarily meant for data
- Numerical data
- Other research data outputs
- Support for common metadata needed for finding or understanding data
  - Supports point, bounding box, and place names for geoLocations
  - Can link metadata record to DataCite record
  - Built in support for many types of relationships to other resources including physical objects
  - Built in support for versioning
- Working to be included in citation indexes, etc.
- Schema and services actively evolving

CrossRef/PILA

- Primarily meant for publications
  - Only registers metadata for works not individual manifestations of a work
  - Schema allows multiple resolution
  - Can register data associated with a work
  - Must also providing linking information about references in the work
- Support for common metadata needed for finding publications or parts thereof
  - Books
  - Journals
  - Conference proceedings
- Well integrated into existing citation metrics, indexing, etc. providers using a pay for query model
- Last schema update in 2004
EZID vs CrossRef/PILA

**EZID**

- EZID supports both DataCite’s DOI’s and ARK’s (lower cost)
- EZID suggests using ARK’s prior to decision to support an object in perpetuity
  - ARK’s can be deleted
  - An ARK can be the suffix of a DOI
  - ARK’s can be used at the granule level using a single registration and a “pass through” suffix
- Annual fee for up to 1 million IDs/yr based on
  - Profit/non-profit status
  - Size/status of organization
- Considering development of single DOI purchase capability

**CrossRef/PILA**

- DOI’s are the only locator supported
- Annual fee based on publishing revenue
- Additional fee for each DOI assigned
- Additional fee if linkage information is not provided for most content within 18 months

Presented to the USGS Digital Object Identifiers Focus Group
By Ruth Duerr, June 12, 2013
CrossRef/PILA Linkage Example
Prepared for Ingest, presented 10/27/09 by R. Duerr

LID590DCL Foundations of Data Curation

EZID vs CrossRef considerations

• Do you have mostly publications or mostly data or both?
• Do you want/need locators prior to making a decision about long-term availability?
• Are citation indexes, citation metrics, or the ability to support full-text access currently important to you?
• Which is more important to you - library or data concepts?
• What kind of metadata do you have about the things you need identifiers for?