

Breakout session 3

Citation and Preservation within the DataONE Framework

Stephen Abrams

California Digital Library

Robert Cook

Oak Ridge National Laboratory

DataONE



Citation and Preservation Breakouts

Nancy Hoebelheinrich
Richard Huffine, USGS
David Miner, SDSC
Paul Bracke, Purdue
Eric Kihn, NGDC

Emily Fort
Mark Servilla
Bob Sandusky
Stephen Richard
Giri Palanisamy
Suzy Allard
Andy Mitchell
Dawn Miller
Dave Rugg
Sky Bristol
Rama Ramapriyan
Alex Joseph

Citation

- Citation Attributes
 - Elements of a citation
 - Exposure in Web of Science
- Citation Benefits (credit authors and enable others to find data)
- Citation Issues
 - Can you trust and use data?
 - How to reference granules?
 - *Reference workflows*
 - Should derived, integrated data products be published?
- What actions the DUG can take with regard to citations and identifiers?
 - Form a subcommittee, prepare a position paper, prepare a set of best practices

Preservation

- Keep the bits safe
- Protect the form, meaning, and behavior of the bits
 - convert data to open format; preserve tools
- Safeguard the guardians
 - federation of D1 Member Nodes
- Preservation Issues
 - Should a data center / MN go out of business, who will preserve the data collection? (Preservation requires effort).

What makes data usable?

It must *exist*, in order to be...

Intellectually, administratively, technically, and legally
described or documented, in order to be...

Preserved over time, in order to be...

Published or shared, in order to be...

Discovered by others, in order to be...

Delivered or downloaded locally in useful form, in order to
be...

Used or transformed

(in which case it must be described, preserved, published, etc.)

Why citation? Why preservation?

Citation is just one component of the larger process of data *access*

- Publication, discovery, delivery, (re)use

Access and preservation are complementary, rather than disparate, activities

- Preservation ensures access *over time*

Citation attributes

Assert unique and persistent data identity

Descriptive metadata useful for purposes of discovery

- Subset of DataONE science metadata or metadata schemas defined/managed locally by MNs
- *DataCite Metadata Schema for the Publication and Citation of Research Data* (July 2011), doi:10.5438/0005 <
http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf>
- Identifier, creator, title, publisher, publication date (*required*)
- Subject, contributor, date, language, resource type, alternative identifier, related identifier, size, format, version, rights, description (*optional*)

Citation attributes

Exposure via well-known disciplinary portals, DataONE CNs, local MN catalogs, and abstracting and indexing (A+I) services

- E.g. Thomson Reuters “Web of Science”

Actionable resolution

Metrics tracking publication and use

- Attribution, accountability, and incentives
- Impact factor (cf. Crossref)

Citation attributes

Standardization of citation form(s), modulo journal and/or publisher requirements

Karimi R, Fisher NS, Folt CL (2010) . Data from: Multielement stoichiometry in aquatic invertebrates: when growth dilution matters. Dryad Digital Repository. doi:10.5061/dryad.1858 <<http://dx.doi.org/10.5061/dryad.1858>>

Bond-Lamberty, B.P. and A.M. Thomson. 2010. A Global Database of Soil Respiration Data, Version 1.0. Data set. Available on-line [<http://daac.ornl.gov>] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/984 <<http://dx.doi.org/10.3334/ORNLDAAC/984>>

Bond-Lamberty, B. and A.M. Thomson. 2010. A global database of soil respiration measurements, Biogeosciences, 7, 1915–1926, 2010. doi:10.5194/bg-7-1915-2010 <www.biogeosciences.net/7/1915/2010/bg-7-1915-2010.pdf>

Citation benefits

Data product citation standards provide basis for increased incentives, recognition, and rewards for scientific data activities

Online digital data enables testing of published hypotheses

- Peer-examination and review of conclusions or analysis based on experimental or observational data

Online digital data allows subsequent users to make new and unforeseen uses and analyses of the same data – either in isolation, or in combination with other datasets.

Citation issues

Disciplinary discovery

- How to support discovery using disciplinary terminology?
- How to support cross-disciplinary discovery?
 - Standards for expressing cross-cutting descriptors, e.g. spatial and temporal coordinates

Branding

- In a distributed network, which copy do you resolve to?
- Are data published (and branded) by DataONE or by individual DataONE MNs?

Citation issues

Granularity

- Scientists lack the necessary constructs and conventions for referring to portions of a database
 - Analogous to the volume and page numbers, or titles, chapters, and sections, used in citing text
- Disciplines may have disparate needs for granular identification
 - What are the differences that need to be addressed distinctly?
- What do you resolve to?
 - Landing page, entire dataset, single dataset component, single table within a component, single element within a table, ...
 - Support for maximal granularity with minimal identifier registration

DOI templates

EZID suffix pass-through

Preservation strategy

Keep the bits safe

Protect the form, meaning, and behavior of the bits

Safeguard the guardians

DataONE Preservation Strategy, PWG workshop, Chicago,
December 5-6, 2010

Preservation strategy

Keep the bits safe

- Redundancy, replication, and heterogeneity
- Audit, error detection, and self-healing
- Good data center practice, e.g. media refresh, disaster recovery/business continuity planning/exercise

Protect the form, meaning, and behavior of the bits

Safeguard the guardians

Preservation strategy

Keep the bits safe

Protect the form, meaning, and behavior of the bits

- Embed data into self-contained and self-describing “objects”
- Know what you have
 - Technical registries and characterization tools
 - Encourage use of non-proprietary, open, transparent, and widely adopted formats, tools, and standards
- Know your rights
 - Ability to hold, make copies/derivatives, transfer to successor
 - Encourage use of CC0
- Cope with obsolescence
 - Migration vs. emulation

Safeguard the guardians

Preservation strategy

Keep the bits safe

Protect the form, meaning, and behavior of the bits

Safeguard the guardians

- *Quis custodiet ipsos custodes?*
- The collectivity of CMs and MNs must guard itself and its constituents
 - Lower technical barriers for membership
 - Self and external audit
 - Organizational sustainability and succession planning

Preservation issues

Preserving bits is (relatively) easy, but is it enough?

- Preserved data may be unusable without preserved tools

Digests are easy for simple octet streams, but more difficult for complex structures

- Canonicalization

Clifford Lynch, “Canonicalization: A fundamental tool to facilitate preservation and management of digital information,” *D-Lib Magazine* 5:9 (September 1999) <<http://www.dlib.org/dlib/september99/09lynch.html>>

- Universal Numeric Fingerprint (UNF)

Micah Altman and Gary King, “A proposed standard for the scholarly citation of qualitative data,” *D-Lib Magazine* 13:3/4 (March/April 2007) <<http://www.dlib.org/dlib/march07/altman/03altman.html>>

Preservation issues

The majority of preservation effort to date has been for cultural heritage material, i.e. text, image, audio/ visual

- Formats used to represent and describe scientific data are less well understood and documented
- Encourage participation by communities of practice
- Microsoft Research/UC Curation center DCXL (Data Curation Excel) project

For more information

DataONE

<http://www.dataone.org/>

DataCite

<https://www.datacite.org/>

Metadata schema

http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

“Data and the Scholarly Record: The Changing Landscape,” Berkeley, 24-25
August 2011

<http://datacite2011.eventbrite.com/>

Stephen Abrams

Stephen.Abrams@ucop.edu

<http://www.cdlib.org/uc3>

Robert Cook

cookrb@ornl.gov

<http://www.ornl.gov/>

Citation and Preservation

Nancy Hoebelheinrich
Richard Huffine, USGS
David Miner, SDSC
Paul Bracke, Purdue
Eric Kihn, NGDC

Emily Fort
Mark Servilla
Bob Sandusky
Stephen Richard
Giri Palanisamy
Suzy Allard
Andy Mitchell
Dawn Miller
Dave Rugg
Sky Bristol
Rama Ramapriyan
Alex Joseph

Citation

- Citation Attributes
 - Elements of a citation
 - Exposure in Web of Science
- Citation Benefits (credit authors and enable others to find data)
- Citation Issues
 - Can you trust and use data?
 - How to reference granules?
 - *Reference workflows*
 - Should derived, integrated data products be published?
- What actions the DUG can take with regard to citations and identifiers?
 - Form a subcommittee, prepare a position paper, prepare a set of best practices

Preservation

- Keep the bits safe
- Protect the form, meaning, and behavior of the bits
 - convert data to open format; preserve tools
- Safeguard the guardians
 - federation of D1 Member Nodes
- Preservation Issues
 - Should a data center / MN go out of business, who will preserve the data collection? (Preservation requires effort).