Persistent Identifiers and Linked Ocean Data

Cynthia L. Chandler, Adam Shepherd, Robert A. Arko and Matthew B. Jones


PIDapalooza Festival Reykjavik, Iceland 9 November 2016
SEA STORIES

THE SPIRIT OF AMERICA IN ACTION

Sea Stories

March 1928

25 CENTS

in Canada 25c

Chester L. Saxby
W. Townend
Don Waters
S.B.H. Hurst
Bill Adams
Lloyd Kay
and others

Google

North Atlantic Fishery

North Atlantic Cheap

www.NexTag.com

Want Deals for North Atlantic? See NexTag Sellers' Lowest Price!
DISTRIBUTED DATA ... GLOBAL SCIENCE

It's not the size of the data...
It's the impact of the science
Distributed Related Resources

- **Funding Source**
  - [Image of National Science Foundation website]
  - [Image of BCO-DMO dataset]
  - [Image of EarthChem Library Record]

- **Data Sets**
  - [Image of BCO-DMO dataset]
  - [Image of EarthChem Library Record]

- **Physical Samples**
  - [Image of sample search]
  - [Image of paper on aerobic microbial respiration]

- **Field Expedition**
  - [Image of field expedition map]
  - [Image of cruise catalog]

- **Publications**
  - [Image of paper on aerobic microbial respiration]
R2R, BCO-DMO, DATAONE

- R2R provides stewardship of routinely-acquired environmental sensor data from U.S. academic research vessels, and a catalog of cruises
- BCO-DMO works in partnership with NSF-funded researchers to help improve access to marine research data some of which is acquired during research cruises
- DataONE is a global network that federates content from many repositories including R2R and BCO-DMO
Sufficient metadata is essential but plain text metadata is not enough. PIDs help, and published as Linked Data is even better to improve discovery and access.
PERSISTENT IDENTIFIERS

Current practices in U.S. ocean science:
- DOIs for Articles
- DOIs for Datasets
- DOIs for Documents
- FundRef codes for Awards
- IGSNs for Samples
- ORCIDs for Persons

(Hanson, 2016; 10.1029/2016EO043183)
Emerging practices in U.S. ocean science:
GRID IDs for Organizations
DOIs for Collections (package of objects)
DOIs for Repositories (via RE3data.org)
DOIs for Expeditions (via eg. R2R, IODP)
DOIs for Networks (via eg. FDSN)

Linking Related Resources via PIDs
Linking Additional Resources (no PIDs yet)

- Educational history
- Research vessel information
- Related research
- Photos of cores
- Instrumentation details
- Related analyses
- Related articles
CONCEPTS IN THE RESEARCH LIFE CYCLE

COVERAGE ACROSS REPOSITORIES
LINKED DATA FOR THE GEOSCIENCES

linkeddata.org     lod-cloud.net
EARTH CUBE GEOLINK PROJECT

http://schema.geolink.org/1.0/base/main
LINKING URIs FOR FEDERATED SEARCH

http://bco-dmo.org/id/person/1234

http://data.nsf.gov/id/researcher/9876

http://schema.geolink.org/1.0/base/main#Person

LINKING URIs FOR FEDERATED SEARCH

http://bco-dmo.org/id/person/1234

http://schema.geolink.org/1.0/base/main#Person

http://bco-dmo.org/id/person/1234

http://data.nsf.gov/id/researcher/9876

surname: Saito
givenName: Makoto

DIFFICULT

surname: Saito
givenName: Mak
LINKING URIs FOR FEDERATED SEARCH

http://bco-dmo.org/id/person/1234

http://data.nsf.gov/id/researcher/9876

http://schema.geolink.org/1.0/base/main#Person

http://schema.geolink.org/1.0/base/main#Identifier

scheme
identifier
LINKING URIs FOR FEDERATED SEARCH

http://bco-dmo.org/id/person/1234

IDENTIFIER
scheme: http://purl.org/spar/datacite/orcid
identifier: 0000-0001-6040-9295

http://data.nsf.gov/id/researcher/9876

IDENTIFIER
scheme: http://purl.org/spar/datacite/orcid
identifier: 0000-0001-6040-9295

surname: Saito
givenName: Makoto

surname: Saito
givenName: Mak
COMPOUNDING THE BENEFITS

Benefits of Linked Data
- Classification
- Expressivity & Relationships
- Linking Resources

Benefits of PIDs
- Identification
- Disambiguation
- Trusted Authority

FESTIVAL TIME

PIDapalooza
Reykjavik, Nov 2016

REGISTRATION IS OPEN!
CONCERN

DOI assignment practices

e.g. assign DOI for an event as well as
collection of results from an event; is there a
recommended practice for citation?

Assign a DOI to the
metadata record for a
sampling event (e.g. cruise)

Assign a DOI to the collection of all
data sets and documents at a repository
that are associated with an event
CHALLENGE

Appropriate relationship between identifiers when encoding provenance?

e.g. dataset isPartOf cruise
dataset isDerivedFrom cruise
cruise isSourceOf dataset

There are multiple reasonable choices, but we need community consensus.

"upstream/downstream" issue
Are there recommended practices for encoding provenance using PIDs.

Datasets pointing to parent event.
Articles reference datasets.
ORCiD points to projects.

Granularity of DOI assignment: DataCite expects DOI PIDs to represent a data set (e.g. with a title), whereas we need PIDs both for that and for components of data sets such as individual files, granules, and records.

e.g. use UUIDs for internal components of data sets, and assign the DOI to the whole data set.

There could be good reason for DOIs to be assigned to more granular components. In general, the PID community has not embraced the need for PIDs at multiple levels of the data hierarchy (cell, record, file, metadata, package).

CHALLENGE

Person IDs: ORCIDs working great for extant people, still challenged on how to assign person identifiers for all of our historical data records (e.g. people who have moved on to other fields)

If a dataset or document isn’t published with a DOI, does it exist?

Will non-DOI’d content be discoverable by emerging networks like CrossCite tool, Scholix, etc?

CONCERN

A DOI is a DOI … or are CrossRef and DataCite data set DOIs different?

One of these things is not like the other … but is there a lossless crosswalk between their metadata schemas?

PROPOSAL

Publish ALL identifiers as Linked Data?

EXEMPLARS:
DataCite DOIs
ORCiD

On deck: IGSN, GRID IDs

CONSIDERATION

Should we publish PIDs for controlled vocabulary terms?

e.g. classes of instruments

Max reference limits:
Publishers limit the number of references (16 or 20 max)
How to cite datasets for a synthesis work that is based on 100s of source datasets?
DOIs issued for multiple copies – different instances of same content

e.g. same data at many repositories

copies of the same data set could appear in many repositories, with different metadata as determined by the repository practices, with each repository assigning a new DOI; in many cases the repositories may not be aware that the data have already been published by another repository

THANK YOU

Research funded by: NSF OCE-1435578 (BCO-DMO), ICER-1440114 (R2R), OCE-1447797 (EarthCube GeoLink)

BCO-DMO: http://bco-dmo.org
info@bco-dmo.org  @BCODMO
cchandler@whoi.edu  @cynDC42
ashepherd@whoi.edu  @ashep_15

R2R: http://rvdata.us
info@rvdata.us
arko@ldeo.columbia.edu