Data Citation: DOI-Enabling GEOSS Discovery and Access

4th GEOSS Science and Technology Stakeholder Workshop
March 27, 2015 -- Norfolk, Virginia

David K Arctur, University of Texas at Austin
Joan Starr, California Digital Library
Stefano Nativi, Enrico Boldrini, and Mattia Santoro,
ESSI Lab, Italy Center for National Research
Robert Arko, Lamont-Doherty Earth Observatory
Abstract

Data citation is enabling much more than just discovery of published resources; it is enabling direct linking of data in a way that enables its use in technical papers, websites, presentations, and other data sets. Furthermore, it represents a contract for permanent availability and access to the referenced data.

From Wikipedia: “A DOI name differs from standard identifiers such as the ISBN. The purpose of an identifier registry is to manage a given collection of identifiers, whereas the primary purpose of the DOI system is to make a collection of identifiers actionable and interoperable.”

There is considerable current discussion on roles and conventions for using DOIs with Earth and space science data. An important aspect of this is the metadata schema used with DOIs. This schema, based on the indecs Content Model, has some overlap with the ISO 19115 content model. DataCite, a consortium of leading research libraries and scientific data centers, has a registry of over 5 million DOIs and their metadata. The DataCite DOI metadata schema has added new features that support the geoscience community, such as a GeoLocation element and the ability to supply extra disciplinespecific metadata. Numerous geoscience data providers routinely publish DOIs using the DataCite schema. By mapping the DOI metadata to ISO 19115, and harvesting DataCite’s DOI registry, the GEO Discovery and Access Broker (DAB) can add an immensely valuable resource to its already significant distributed catalogs.

This presentation reviews the current discussion of DOI usage in geoscience research, and the role that GEOSS can now play in connecting this with a broader community.
Data Citation: DOI-Enabling GEOSS

Outline

• Introduction to Digital Object Identifiers and DataCite
  (Joan Starr)

• Initial registration of DataCite.org’s DOI registry with GEO Discovery and Access Broker (DAB)
  (Stefano Nativi, Enrico Boldrini, Mattia Santoro)

• Questions raised, next steps
  (All)
DataCite Metadata and ISO 19115

Joan Starr
California Digital Library
What is a DOI?

Mycological Progress
August 2012, Volume 11, Issue 3, pp 827-833

Lepidostroma vilgalysii, a new basidiolichen from the New World

Brendan P. Hodkinson, Jessie K. Uehling, Matthew E. Smith


Within this Article
» Introduction
» Materials and methods
» Results
» The new species
» Discussion
» References


Within this Article
» Introduction
» Materials and methods
» Results
» The new species
» Discussion
» References


Within this Article
» Introduction
» Materials and methods
» Results
» The new species
» Discussion
» References

Lepidostroma vilgalysii, a new basidiolichen from the New World

Brendan P. Hodkinson, Jessie K. Uehling, Matthew E. Smith


Within this Article
» Introduction
» Materials and methods
» Results
» The new species
» Discussion
» References
What is a DOI?

*What you see:* alphanumerical string (*never changes*)

*Associated with:* location of object (*such as a URL*)

*And:* who, what, when, etc (*i.e. metadata*)


DOI example

**string:** doi:10.9999/FK40K2GTV

**html version:** [http://dx.doi.org/10.9999/FK40K2GTV](http://dx.doi.org/10.9999/FK40K2GTV)

**location:** [http://www.bologna.edu/biology/xfg/123.xls](http://www.bologna.edu/biology/xfg/123.xls)

**metadata**

- **creator:** Dr. Felix Kottor
- **title:** Data for chromosomal study of catfish (Ictalurus punctatus)
- **publisher:** University of Bologna
- **date:** 8/31/2012
DOI example

string: doi:10.9999/FK40K2GTV

html version: http://dx.doi.org/10.9999/FK40K2GTV

location: http://www.state.edu/ecology/783sdr/123.xls

metadata

creator: Dr. Felix Kottor

title: Data for chromosomal study of catfish (Ictalurus punctatus)

publisher: Dryad Data Repository

date: 10/01/2013
Why are DOIs important?

The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Please try the following:

- If you typed the page address in the Address bar, make sure that it is spelled correctly.
- Open the [httpd.apache.org](http://httpd.apache.org) home page, and then look for links to the information you want.
- Click the Back button to try another link.
- Click Search to look for information on the Internet.

HTTP 404 - File not found
Internet Explorer
Why are DOIs important?

Example:

Allow readers to **find** data products
Get **credit** for data and publications
Promote **reproducibility**
Better measure of research **impact**
DataCite
Creating a global citation framework for data
DataCite Services

1. DOIs for data!
2. Local service & support
3. Usage stats
4. Citation formatter
5. Content negotiation
6. Metadata search
7. OAI provider
8. DataCite-to-ORCID hookup
DataCite Metadata

• Mandatory elements support data citation
• Recommended elements support indexing, discovery
• Optional elements support additional description, access, management, etc.

• Give your feedback, questions, input here: bit.ly/1I6F3Im
For more information

- DataCite metadata scheme
- DataCite metadata search
- All about DataCite
Data Citation: DOI-Enabling GEOSS

Outline

• Introduction to Digital Object Identifiers and DataCite (Joan Starr)

• Initial registration of DataCite.org’s DOI registry with GEO Discovery and Access Broker (DAB) (Enrico Boldrini, Mattia Santoro, Stefano Nativi)

• Questions raised, next steps (All)
DataCite Archive Interoperability

Enrico Boldrini, Mattia Santoro and Stefano Nativi

CNR-IIA, UOS di Firenze
Brokering DataCite archive

http://oai.datacite.org/oai?

OAI-PMH service with DataCite metadata

Mapping metadata to ISO 19115
DataCite results from the GI-portal

- Partial harvest of DataCite catalog (2588 records)
- Full catalog: > 5 million records
Query by text « cmip5 »

<table>
<thead>
<tr>
<th>Access/Use Constraints</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CMIP5 simulations of the Max Planck Institute for Meteorology (MPI-M) based on the MPI-ESM-LR model: The amipFuture experiment, served by ESGF</td>
</tr>
<tr>
<td></td>
<td>CMIP5 simulations of the Max Planck Institute for Meteorology (MPI-M) based on the MPI-ESM-MR model: The amipFuture experiment, served by ESGF</td>
</tr>
<tr>
<td></td>
<td>cmip5 output1 NCC NorESM1-M sstClim4xCO2, served by ESGF</td>
</tr>
<tr>
<td></td>
<td>NOAA GFDL GFDL-ESM2G, esmrcp85 experiment output for CMIP5 AR5, served by ESGF</td>
</tr>
<tr>
<td></td>
<td>NOAA GFDL GFDL-ESM2G, rcp45 experiment output for CMIP5 AR5, served by ESGF</td>
</tr>
<tr>
<td></td>
<td>NOAA GFDL GFDL-ESM2G, esmControl experiment output for CMIP5 AR5, served by ESGF</td>
</tr>
</tbody>
</table>
NOAA GFDL GFDL-ESM2G, esmrcp85 experiment output for CMIP5 AR5, served by ESGF

ISO 19115 Overview (see raw metadata)
GEOSS Search for « esm2g AND esmrcp85 »

<table>
<thead>
<tr>
<th>Record</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>File identifier</td>
<td>de.dkrz.wdcc.iso3205549</td>
</tr>
<tr>
<td>Hierarchy level</td>
<td>series</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Identification Information</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>cmip5 output1 NOAA-GFDL GFDL-ESM2G esmrcp85</td>
</tr>
<tr>
<td>Creation date</td>
<td>2014-04-03</td>
</tr>
</tbody>
</table>
How can we know these are the same datasets?
Data Citation: DOI-Enabling GEOSS

Outline

• Introduction to Digital Object Identifiers and DataCite
  (Joan Starr)

• Initial registration of DataCite.org’s DOI registry with GEO Discovery and Access Broker (DAB)
  (Stefano Nativi, Enrico Boldrini, Mattia Santoro)

• Questions raised, next steps
  (All)
Questions raised, next steps...

- Key benefit of supporting DOI within GEOSS:
  - Enabling permanent, unique, resolvable links to data within documents

- Questions about metadata mapping between DataCite and ISO 19115
  - Handling of fields optional for DOI but required for ISO 19115 by INSPIRE

- Possible next steps
# DataCite to ISO 19115 Mapping

<table>
<thead>
<tr>
<th>DataCite Requirements</th>
<th>ISO 19115 support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier</td>
<td>identificationInfo[1]/<em>/citation/</em>/identifier[1]</td>
</tr>
<tr>
<td>Creator</td>
<td>identificationInfo[1]/*/pointOfContact, role=originator</td>
</tr>
<tr>
<td>Title</td>
<td>identificationInfo[1]/<em>/citation/</em>/title</td>
</tr>
<tr>
<td>Publisher</td>
<td>identificationInfo[1]/*/pointOfContact, role=publisher</td>
</tr>
<tr>
<td>PublicationYear</td>
<td>identificationInfo[1]/<em>/citation/</em>/date, dateType=publication</td>
</tr>
</tbody>
</table>

**EXAMPLE:** Creator (PublicationYear): Title. Publisher. Identifier

A. H. Allen (1916): Juvenile Leucosticte on rocks. Museum of Vertebrate Zoology. [http://dx.doi.org/10.7299/X7PV6HQ0](http://dx.doi.org/10.7299/X7PV6HQ0)
## Mapping of DataCite Mandatory Properties

<table>
<thead>
<tr>
<th>DataCite ID</th>
<th>DataCite Property</th>
<th>ISO 19115 Property (XPath)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identifier</td>
<td>identificationInfo[1]/<em>/citation/</em>/identifier[1]</td>
</tr>
<tr>
<td>2</td>
<td>Creator</td>
<td>identificationInfo[1]/*/pointOfContact, role=originator</td>
</tr>
<tr>
<td>3</td>
<td>Title</td>
<td>identificationInfo[1]/<em>/citation/</em>/title</td>
</tr>
<tr>
<td>4</td>
<td>Publisher</td>
<td>identificationInfo[1]/*/pointOfContact, role=publisher</td>
</tr>
<tr>
<td>5</td>
<td>Publication Year</td>
<td>identificationInfo[1]/<em>/citation/</em>/date, dateType=publication</td>
</tr>
</tbody>
</table>
## Mapping of DataCite Recommended and Optional Properties

<table>
<thead>
<tr>
<th>DataCite ID</th>
<th>DataCite Property</th>
<th>ISO 19115 Property (XPath)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Subject</td>
<td>identificationInfo[1]/<em>/descriptiveKeywords/</em>/</td>
</tr>
<tr>
<td>7</td>
<td>Contributor</td>
<td>identificationInfo[1]/*/pointOfContact, various roles</td>
</tr>
<tr>
<td>8</td>
<td>Date</td>
<td>identificationInfo[1]/<em>/citation/</em>/date, various date types</td>
</tr>
<tr>
<td>9</td>
<td>Language</td>
<td>identificationInfo[1]/*/language</td>
</tr>
<tr>
<td>10</td>
<td>ResourceType</td>
<td>hierarchyLevel</td>
</tr>
<tr>
<td>11</td>
<td>AlternateIdentifier</td>
<td>identificationInfo[1]/<em>/citation/</em>/identifier[2]</td>
</tr>
<tr>
<td>12</td>
<td>RelatedIdentifier</td>
<td>aggregateDataSetIdentifier</td>
</tr>
</tbody>
</table>
## Mapping of DataCite Recommended and Optional Properties

<table>
<thead>
<tr>
<th>DataCite ID</th>
<th>DataCite Property</th>
<th>ISO 19115 Property (XPath)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Size</td>
<td>distributionInfo[1]/transferOptions/*/transferSize</td>
</tr>
<tr>
<td>14</td>
<td>Format</td>
<td>distributionInfo[1]/distributionFormat/*/name</td>
</tr>
<tr>
<td>15</td>
<td>Version</td>
<td>distributionInfo[1]/distributionFormat/*/version</td>
</tr>
<tr>
<td>16</td>
<td>Rights</td>
<td>identificationInfo[1]/resourceConstraints/otherConstraints</td>
</tr>
<tr>
<td>17</td>
<td>Description</td>
<td>identificationInfo[1]/abstract</td>
</tr>
<tr>
<td>18</td>
<td>GeoLocation</td>
<td>identificationInfo[1]/extent/geographicElement</td>
</tr>
</tbody>
</table>
Almost all (18) DataCite elements can be easily mapped to ISO 19115.

However ISO 19115 (and especially INSPIRE) require additional mandatory elements that are **missing** or **optional** in DataCite. E.g.

- Metadata fileIdentifier
- Metadata date stamp
- Organisation email
- Geographic bounding box
- Metadata language
- Lineage
- Abstract
- GEMET keywords
- Conformity to INSPIRE
- Conditions/limitations on access and use
## Suggested ISO 19115 to DataCite Mapping

<table>
<thead>
<tr>
<th>ISO 19115 Requirement</th>
<th>DataCite support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata file identifier</td>
<td>use relatedIdentifier, relationType=HasMetadata to point to associated, additional metadata</td>
</tr>
<tr>
<td>Metadata date stamp</td>
<td>access via OAI-PMH</td>
</tr>
<tr>
<td>Organization email</td>
<td>use nameIdentifiers (ORCID, FundRef, ISNI, etc.) to point to additional information for creators and contributors</td>
</tr>
<tr>
<td>Geographic bounding box</td>
<td>use GeoLocation and geoLocationBox</td>
</tr>
<tr>
<td>Metadata language</td>
<td>along with HasMetadata, use relatedMetadataScheme, schemeType and schemeURI</td>
</tr>
<tr>
<td>Lineage</td>
<td>use relatedIdentifier with various relationTypes to describe provenance (isSubsetOf, isPartOf, isNewVersionOf, etc.)</td>
</tr>
<tr>
<td>Abstract</td>
<td>use Description; descriptionType=Abstract</td>
</tr>
<tr>
<td>GEMET keywords</td>
<td>use Subject; subjectScheme=GEMET</td>
</tr>
<tr>
<td>Conformity to INSPIRE</td>
<td>use relatedIdentifier, relationType=HasMetadata to point to associated, additional metadata</td>
</tr>
<tr>
<td>Conditions on access/use</td>
<td>use Rights, which is repeating</td>
</tr>
</tbody>
</table>
DataCite recommended approach

• Describe relationships between research objects.
• Use name and organizational identifiers whenever available.
• Take advantage of DataCite’s OAI-PMH service to retrieve GEOSS metadata in desired format (e.g. RDF).
Next steps...

• How will GEO architect support for DOI linking and other uses within GEOSS?
  o DOI as a search qualifier
  o Generate/assign DOIs for datasets with no other ID?

• How should GEO work with DataCite to:
  o Ensure that the design of DOI-integration in GEOSS is consistent with community practice
  o Consider GEOSS-INSPIRE recommended profile for DOI metadata

• Coordination through DataCite membership? Research Data Alliance? Force11? EarthCube? …?
Thank you

David Arctur
david.arctur AT utexas.edu