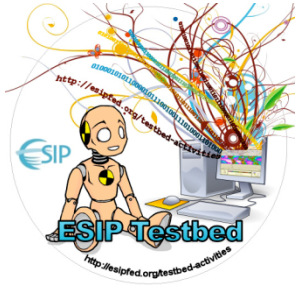
The Realities of Implementing ID Schemes for Data Objects:

An ESIP Products & Services Testbed / Data Stewardship Committee Project



ESIP Federation Summer Meeting 2012

Madison, WI 17 – 20 July 2012

By Nancy J. Hoebelheinrich, Knowledge Motifs LLC & Greg Janée, UC Santa Barbara

Knowledge Motifs LLC

Mapping sensible data relationships

Overview – what's to be covered

- Background & Problem Statement
- Process & Findings for 8 ID schemes
- Conclusions
- Implications / Issues for further discussions
- Questions / Discussion

Identifiers for Data Objects

Part 1 Complete

Part 1 – Use Cases

- ***Unique Identification***: To uniquely and unambiguously identify a particular piece of data, no matter which copy a user has
- ***Unique Location***: To locate an authoritative copy of the data no matter where they are currently held
- ***Citable Location***: To identify cited data
- ***Scientifically Unique Identification***: To be able to tell that two data instances contain the same information even if the formats are different

Eight ID Schemes Assessed

- DOI
- ARK
- UUID
- XRI
- OID
- Handles
- PURL
- LSID
- [URI, URN, URL]

Ruth E. Duerr, Robert R. Downs, Curt Tilmes, Bruce Barkstrom, W. Christopher Lenhardt, Joseph Glassy, Luis E. Bermudez and Peter Slaughter. *On the utility of identification schemes for digital earth science data: an assessment and recommendations*, Earth Science Informatics, Vol. 4, Num. 3, 139-160, 2011, doi:10.1007/s12145-011-0083-6.

Identifiers for Data Objects

Part 1 Complete

Categories of ID Characteristics	**Selected Questions asked of each ID Scheme**
Technical Value	• How scalable is it to very large numbers of objects?
User Value	• Will publishers allow it in a citation?
Archive Value	• How maintainable is the identification scheme when data migrates from one archive to another (or even from one location in an archive to another)?

http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Identifiers/Table

Identifiers for Data Objects

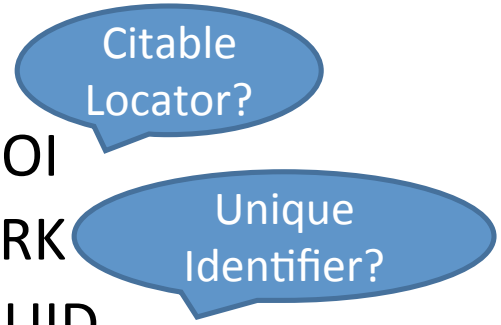
Part 2 - in process

Part 2 – ID Testbed

Premise:

Recommendations ***need to be tested by implementation*** to address any operational issues that might change the recommendations

Schemes Recommended



- DOI
- ARK
- UUID
- XRI
- OID
- Handles
- PURL
- LSID
- [URI, URN, URL]

Data sets used: two types

Image Collection

- The Glacier Photo Collection from the National Snow and Ice Data Center
  - A photographic image collection of glaciers in various image formats ranging from JPEG to TIFF (http://nsidc.org/data/glacier_photo/)

Numerical Time Series

- MODIS sensor data from NASA's Earth Observing System (EOS) Aqua and Terra satellites
  - A subset of Level-2 and Level-3 data products representing snow cover and sea ice http://nsidc.org/data/modis/.

Use Case Requirements Translated to Testable Questions

Requirements by Use Case per Duerr et al Identifier Paper

	A	B	C	D	E	F
1	**Paper Requirement**	**Means for testing in Testbed environment**	**Unique ID use case?**	**Unique locator use case?**	**Citable ID use case?**	**Scientifically unique Use Case?**
2	Location "independence" **	Is a property of an identifier. Preferably, would be embedded in the data object or be computed, like a checksum. Did not test.	x			x
3	Location "invariance" **	Is a property of an identifier. Would be able to ascertain if the DO could be found regardless of its current location. Did not test.		x	x	
4	Generated at time of DO creation	Restrictions on when DO can be created? Who can create? Can be created in the field? For testbed, DO already created. N.A. except for process of assignment of IDs.	x			x
5	Can be created after DO is considered to be permanently available	Who can create? Data producer or data archive? Can be created in the field? For testbed, DO already created. N.A. except for process of assignment of IDs.		x	x	x
6	ID necessarily placed within or carried along with DO	Not practical to test; is a practice based on usage.	x			x
	Referenced with	Is descriptive MD associated with the DO? If so, how	x			x

http://wiki.esipfed.org/images/b/bf/OrigReqs.pdf

Operational Questions Asked

Categories of Operational Issues	Selected Questions asked of each ID Scheme
ID Assignment / Maintenance Issues	• What is relationship with URI? (Addresses unique identifier capability, and interoperability)
Discovery Issues	• Is it possible or recommended that descriptive metadata be associated with the ID? If so, how maintained?
Archival Issues	• How is the association between the ID and the resource maintained when transferred from one archive or repository to another (e.g., embedded within object?)

Full list of ?'s: http://wiki.esipfed.org/images/8/82/OpsQuestionsAsked.pdf

Answers to date: http://wiki.esipfed.org/images/9/92/SummaryTestingAnswers_draft1.pdf

DOIs

doi:10.1007/11551362_27

- Global, distributed infrastructure for registration and resolution
- Setup
  - Need allocator
    - We used IDF → DataCite → CDL (EZID)
  - Sign contract (with user responsibilities), annual subscription payment
    - No more per-identifier charges
  - Set up group & user accounts, prefixes

DOIs

- Created data set-level identifiers
  - Glacier photos: doi:10.5060/D4RN35SD
  - MODIS: doi:10.5060/D4CC0XMZ
  - Registration takes ~3 seconds
    - With concurrent threads can get sub-second throughput
- Opaque identifiers created ("minted") by EZID
  - Custom identifiers possible
- Resolvable URLs
  - http://dx.doi.org/10.5060/D4CC0XMZ
- EZID metadata records viewable/editable at
  - http://n2t.net/ezid/id/doi:10.5060/D4CC0XMZ

EZID

long-term identifiers made easy

Login | Contact Us

Home **Manage IDs** Create IDs Lookup ID Demo

Identifier Details

Identifier: doi:10.5060/D4CC0XMZ Get link

About the identified object

Location (URL):	http://nsidc.org/data/myd10cmV5.html
Creators:	Dorothy K. Hall; Vince V. Salomonson; George A. Riggs
Title:	MODIS/Aqua Snow Cover Monthly L3 Global 0.05Deg CMG, Version 5
Publisher:	National Snow and Ice Data Center (NSIDC)
Publication year:	2006
Subjects:	Cryosphere [GCMD]; Snow/Ice [GCMD]; Snow Cover [GCMD]
Resource type:	Dataset
Formats:	HDF-EOS
Version:	V005
Rights:	No access or use constraints.
Description [Abstract]:	The MODIS/Aqua Snow Cover Monthly L3 Global 0.05Deg CMG (MYD10CM) data set, new for Version 5 (V005), contains snow cover and Quality Assessment (QA) data in Hierarchical Data Format- Earth Observing System (HDF-EOS) format, and

DOIs

- Granule-level identifiers
  - To create unique locators, granules must have URLs
  - Easily true for glacier photos
  - Only quasi-true for MODIS data set
    - Multiple URLs/granule tied together through NSIDC data access system
  - DataCite requires “landing pages” URLs

DOIs

- Granule-level identifier naming approaches
  - Individual, unrelated (generated) names
    - + supports complete freedom of granule movement
    - - requires maintaining database of mappings
  - Individual names: base identifier + local granule name
    - doi:10.5060/D4RN35SD/baird1929090101
    - + no database required (mapping is functional)
    - - can't rename granules
  - One, base identifier with partial redirect
    - doi:10.5060/D4RN35SD/local → http://nsidc.org/cgi-bin/gpd_deliver_jpg.pl?local
    - + only one identifier to manage
    - - granules must be managed as a group
    - Not supported by DOIs (PURLs yes, Handles 7+, ARKs someday)

DOIs

- Citation metadata required
  - DataCite schema required
    - Publication-oriented
    - Mandatory: title, creator, publisher, publication year, resource type (controlled vocabulary)
  - EZID starting to provide mappings to DataCite
- Publishers have expressed interest in harvesting DataCite, EZID metadata
- Ergo, DOIs support citable locators

ARKs

ark:/13030/m5k9386w

- Similar to DOIs
  - Have both URI and URL syntaxes
  - Hierarchical decomposition
  - Emphasis on opaque components
  - Global infrastructure
  - Central, HTTP/URL-based resolver (http://n2t.net, "name-to-thing," in the case of ARKs)
  - Ability to associate metadata with identifiers
  - Support citable locators
- Creation identical through EZID

ARKs

	DOIs	ARKs
Support	Distributed, redundant ✓	CDL only (but expanding)
Cost	More expensive (but no longer per-identifier)	Cheaper ✓
Resolution	Must resolve to landing page	Unrestricted; partial resolution forthcoming ✓
Metadata	Citation metadata required; DataCite schema only	Optional, but if present supports citation; multiple schemas
Scope	Published-like things	Anything
Acceptance	Accepted by publishers ✓	Possibly accepted

PURLs

http://purl.org/OCLC/RSPD

- PURL website manages identifiers, users, groups, domains
- First: create user account
- Second: create domain
  - http://purl.org/domain/identifier…
  - Hard!
    - To create opaque domain, that is

PURLs

- Created exact and partial-redirect PURLs
  - http://purl.org/5060D4/RN35SD/okpilak2004080501
  - http://purl.org/5060D4/glacier_photos/{photo_id}
  - Registration takes ~1 second
- Batch API
  - Missing and incorrect documentation
  - Working code included in report
- No associated metadata
  - Less support for long-term maintenance, citation

UUIDs

75c47164-d1a9-11e1-b357-0025bce7cc84

- Supported by all major programming languages
- Suitable (perfect?) as unique identifiers
- Any other use will require significant effort

OIDs

1.3.6.1.4.1.1466.115.121.1.26

- OIDs are hierarchical; each node in tree is authority over subnodes
- No global infrastructure
  - No implementation checks on collision, ownership
- Registration methods
  - RFC process
  - IANA Private Enterprise Number
    - Amounts to embedding owner name in identifier
    - Prevents changes of ownership

OIDs

- UUID from OID Repository
  - Fill out registration form
  - Get back
    2.25.699328204197855217309966167090149307 15
  - Human-mediated

OIDs

- Asked for “NSIDC Glacier Photo Collection”
  - “Note that the description has been changed to NSIDC. … The identifier has also been changed to ‘nsidc’.”
- Asked for “MODIS Snow Cover Climate Modeling Grid (CMG), Monthly Level 3 Global Product at 0.05Deg Resolution”
  - “The description does not makes sense. Such an OID can only be assigned to a company or an international project.”

OIDs

- Tried creating OID for NSIDC
  - And then manage ownership of subnodes
  - Never received registration requests

XRIs

@nsidc.org+dataset*newName!(doi:10.1234/…)

- Contentious
  - Rejected for standardization at W3C's urging
- No activity since 2008
- Resolver (xri.net) redirects to inames.net
- I-names: DNS replacement
  - Low/nonexistent adoption
  - Major I-names (Google, Facebook) are available
- For persistence, need to use I-numbers anyway
  - No benefit over DOIs

Handles

hdl:4263537/5555

- Handle System defined by RFCs
  - But in reality, defined by CNRI Java software
- Local identifiers
  - Run local server
- Global identifiers
  - Run local server
  - Obtain prefix(es) from CNRI (annual fee)
  - Register local server with global Handle system

Handles

- Test:
  - Installed server on Amazon EC2
  - Used Java client in UI & batch modes
  - Registrations took 2-3 seconds
- Metadata
  - Supported by Handle System
  - Not used
  - Ergo, Handles perhaps less suitable as citable locator

LSIDs

urn:lsid:ncbi.nlm.nig.gov:GenBank:T48601:2

- LSID = URN combining local identifier with domain name of owner
  - urn:lsid:ncbi.nlm.nig.gov:GenBank:T48601:2
- No global infrastructure
  - Resolution through DNS, domain names
- Intended to be layered on existing databases
  - Multiple implementations
  - Multiple packages (Java, Perl, .NET)
  - Validation program

LSIDs

- Incorporation of domain names inhibits use as unique locators
- Low adoption rate
- “The experiment [in LSIDs] has not gone well, in the sense that the support for LSIDs has waned over the years, and the only active use of LSIDs comes from a small community in biodiversity informatics. From my perspective, the LSID spec is all but abandoned.”

Conclusions

- Most suitable: DOIs, ARKs, Handles
  - In terms of ease of use, support, adoption, scalability
  - As unique locators
  - DOIs and ARKs support citable locators
    - Publishers have expressed interest in harvesting identifier metadata
  - Handles require more investment
    - Local, dedicated server

Conclusions (cont'd)

- UUIDs: best as unique identifiers
- LSIDs
  - Possibly suitable as unique identifiers, but unsuitable as unique locators due to reliance on domain names
  - Low adoption
- PURLs
  - No support for creating opaque identifiers
  - Poor API support
- Least suitable: XRIs, OIDs
  - XRIs not operational
  - OIDs targeted at other resource types

Next Steps for Testbed

- Finish Summary Spreadsheet with questions answered for each ID scheme
- Finish *draft* Report on Findings
- Vet Report & Summary with ESIP
- Finalize Report & publish?

NEXT?
Identifiers for non-Data Objects

What about IDs for "non-data" objects?

- What kinds of objects need identifiers? (Researcher IDs?)
- Who else is exploring these issues? DataOne?
- Can re-use of data sets be facilitated? (IDs for algorithms, instrument sensors, subsets of data sets used for specific studies)

Some of the questions being asked:

- What is importance for long-term discovery?
- What is relevance to long-term archiving?
- What is importance for data stewardship?