Preliminary Identifiers' Testbed Results: ESIP Data Preservation & Stewardship Cluster

Nancy Hoebelheinrich

Knowledge Motifs LLC

Review: Purpose of Testbed

- Wanted! *Reliable* & *Accurate Citations* that facilitate ability to:
  - FIND
  - IDENTIFY
  - LOCATE
  - OBTAIN
- data sets & their granules (component files or sub-files)

Review: Purpose of Testbed

- Wanted! *Reliable* & *Accurate Citations* to data sets & their granules
  - Via traditional publication sources
  - For non-traditional uses such as Google mashups and other web-accessible paths

Key first step -- ***IDENTIFIERS***

- What about ‘em? They should be:
  - Unique (per data set & granule)
  - Persistent (valid as long as the science that it represents is valid)
  - Actionable (resolve & locate referenced data set)

What means *Unique* Identifiers?

- Capability of returning identifying characteristic about data set that will not conflict with others w/in & w/out system or service
- Capability of identifying discrete granules of a data set that will not be confused with others in the data set
- Hook by which internal and external services or systems can find a data set & its granules

What means *Persistent* Identifiers?

- Capability of:
  - returning identifying characteristic about data set & its discrete components that will remain viable over time
  - Distinguishing among data set versions
  - Locating a data set and its components within internal / external services or systems regardless of changes to the services or systems

What means *Actionable* Identifiers?

- Capability of resolving the identify of a data set & its discrete granules for the ultimate purpose of achieving location
- Return of the resource should occur regardless of the physical location of the resource(s)
- Or, metadata should be returned to point to the location of the resource(s)

Background of chosen dataset for first phase

- Glacier Photograph Collection, National Snow & Ice Data Center
  - Collection of ~ 13 K digitized photos of glaciers taken over a span of ~ 30 years
  - Descriptive MD at data set level
- Date span of 1870 – 2010
- File formats include jpgs for thumbnails, tiffs for high resolution
- Open, but slow growth resulting from voluntary submissions
- Need IDs for each image in collection

Initial Candidate Identifier schemes considered

- DOI
- UUID
- PURL
- OID

- **Which schemes work together?**

- Others to include:
  - ARK
  - XRI
  - LSID
  - Handles

Methodology planned

- Investigate advantages and disadvantages of each identifier scheme from POV of documentation
- Assign identifiers from each scheme to the data set, and each file within the data set & note the problems associated with the assignment of identifier
- Report to group, evaluate & iterate with combination of ID schemes, if needed

Questions for each ID scheme

- What is the unit of identification?
- Can (separate) IDs apply to the data set, to each file & sub-files within the data set?
- How do IDs apply to versions and / or copies of data sets & files/sub-files within the data set?
- What is mechanism for declaring relationships among data set, files /sub-files and versions / copies?

Questions for each ID scheme, cont.,

- What are the requirements for naming or registration entities assigning IDs?
- What metadata, if any is associated with the ID?
- What is / can be the relationship of the ID to other IDs?
- What are costs, if any, associated with use of ID scheme?

Answers to ?'s for DOIs

- What is the unit of identification?
- Can (separate) IDs apply to the data set, to each file & sub-files within the data set?
- How do IDs apply to versions and / or copies of data sets & files/sub-files within the data set?

- "any referent involved in intellectual property transaction" determined to be a "first class content object"
- Yes, decided by registrant
- Use of *DOI system* or framework which includes numbering, description, resolution & policies (DOI Handbook, Edition 4.4.1, October 2007, p. 17)

Answers to ?'s for DOIs, cont.,

- What is mechanism for declaring relationships among data set, files /sub-files and versions / copies?
  - Relationships declared in associated metadata based on *indecs* data dictionary
- Appears to be via typing:
  - e.g., structuralType (closed list values = abstraction, performance, digital, physical)
  - resourceType (not closed list, so could be extended), now closer to book & jrnl publishing genres, e.g., Journal Issue, Journal Abstract
  - Use of legacy identifiers

Answers to ?'s for DOIs, cont.,

- What are the requirements for naming or registration entities assigning IDs?

- DOI Registration Agency (RA) requirements:

  - 1. An RA must be capable of producing a Kernel Metadata Declaration (KMD) for each DOI name issued.
  - 2. Metadata exchanged between RAs supporting DOI System services should be exchanged using an agreed DOI System Referent Metadata Declaration (RMD) for the Referent or Service type.
  - 3. Proprietary terms (data elements and values) used by RAs in KMD & RMD Declarations should be registered in the IDF's data dictionary.

Answers to ?'s for DOIs, cont.,

- What metadata, if any is associated with the ID?
  - Some associated MD required, other optional but strongly recommended for the following purposes related to MD interoperability:
    - 1. To ensure that metadata held by different RAs is *not fundamentally inconsistent*,
      - Via use of Kernel MD
    - 2. To ensure that an *efficient and extensible means of interchange* exists for transporting metadata between RAs (and in future other service providers).
      - Via use of interchange provisions of Referrent MD & indecs data dictionary

Answers to ?'s for DOIs, cont.,

- What is the relationship of the ID to other standard IDs?
- Use legacy identifiers considered "normal" practice
  - Often included as suffix section of DOI ID
  - But, relation described via associated metadata

Answers to ?'s for DOIs, cont.,

- What are costs, if any, associated with use of ID scheme?
  - Expectation of a few cents charged by Registration Agency (RA) upon assignment of DOI
- Designed to be a self-sustaining system (*not* for profit)
- Funds help RA maintain DOI infrastructure incl MD stores

Implementation problems TBD

- In process of identifying registration agency to be used
- Identifying costs for assigning DOIs to Glacier Photo data set and granules
- Identifying mechanism for assigning DOIs to each file using existing MD – are there tools or APIs available?

Implementation problems TBD, cont.,

- Documenting issues that arise
- Testing the IDs in citations, including web-based to see that data set / granules can be identified and located

On to the next ID scheme: UUIDs…

- Plan to report in full by next ESIP meeting, if possible
- Include recommendations for ID schemes to use for ESIP data sets

Questions?

- Nancy Hoebelheinrich, Knowledge Motifs LLC
  - nhoebel@kmotifs.com
- Ruth Duerr, National Snow & Ice Data Center
  - rduerr@nsidc.org