Why USGS Metadata Records Need Persistent Identifiers

LISA ZOLLY
USGS SCIENCE ANALYTICS & SYNTHESIS PROGRAM
CDI METADATA REVIEWERS COP MEETING, 3 FEBRUARY 2020
How does my metadata record get published to catalogs?

**yes, the bird’s name is correctly spelled ‘emu,’ but isn’t that face way more fun than a stock WAF icon?
Without a unique identifier embedded in each record, we can’t know if a harvested record is new, updated, or unchanged. We therefore have to use a dump and replace harvest process weekly to capture new items, revisions, and deletions.

Titles are poor indicators of uniqueness.

DOIs won’t work as unique IDs for metadata:

- ~50% of our metadata lack DOIs (pre-October 2016 data releases)
- Not always 1 DOI : 1 dataset/metadata record. Sometimes 1 DOI : multiple datasets/metadata records
Weekly dump and replace is inefficient and sometimes problematic

- Catalog now at 20,000+ records
- Job takes several hours to run
- Unexpected events can halt harvesting process, resulting in loss of records
  - Network issues
  - WAF inaccessible
  - ScienceBase feed or server interruptions
- Can’t account for quantitative or quantitative changes from week to week
Downstream impacts: WRET CMS

- When harvest is incomplete, WRET CMS may not get new records
- WRET requires a unique identifier for the CMS; they are using the DataCite DOI to meet this requirement
- Consequences:
  - Only metadata containing a 10.5066 DataCite DOI in the primary onlink can be harvested into the CMS
  - For centers/programs assigning DOIs at project/event level to multiple datasets/metadata records, 1:many relationship doesn’t work in CMS
Downstream impacts: other catalogs

- Downstream catalogs all practice incremental harvesting, and rely on unique identifiers to determine:
  - Whether a record is brand new
  - Whether a record is an update to an existing record, and should replace the older version
  - Whether a record has been withdrawn from SDC and should be removed from the downstream catalog

- At the moment, these catalogs are forced to use titles as unique identifiers for USGS metadata. This has produced problems.
  - Titles can change when metadata are versioned
  - Some titles are not unique in our collections
Conclusions

- SDC needs to move to incremental harvest approach to track, manage, and provide metrics for items in growing catalog (growing 20-25% annually).

- In order to do this, SDC will begin this year to require unique persistent identifiers in each metadata record submitted.
Other drivers

- WRET is asking for SDC to provide records with unique PIDs to make available to Drupal sites the full set of a center/program’s metadata collection.

- Data.gov and GeoPlatform will begin requiring provision of unique PIDs in metadata records in the near future to ensure their catalogs are accurate and up to date.

- DataONE needs unique PIDs in our metadata to harvest them correctly in their new system.

- Google Data Search requires unique PIDs in our records; currently, ORNL is faking them for us in the schema.org they are creating in our ‘slim XML’ versions they provide to GDS. We have to move these to USGS infrastructure this year.
How will unique PIDs be generated for metadata?

- These unique PIDs are **not** DataCite DOIs. Continue to assign a DataCite DOI to the data release. The PID is for the metadata record itself.
- SAS is currently developing a Persistent Identifier (PID) Tool that will be used for this purpose.
  - Tool will allow you to generate a unique PID for your metadata record within 60-90 seconds. This is a very simple record!
    - Login → indicate whether PID is for metadata or image → provide title of the item → click Submit to generate the PID to be used in your metadata record → BOOM!
  - PID Tool will:
    - Have a simple user interface for individual generation and assignment (onesies)
    - Have an API for batch generation and assignment by super user WAF owners
    - Be integrated into the OME (auto-created for new records, assigned for uploaded records lacking the PID)
Where do I insert the PID in my CSDGM metadata record?

- CSDGM has no dedicated element for a unique identifier for the metadata record itself.
- We can’t insert a PID in an XML header of the file, because downstream catalogs rely on crosswalks and will ignore header content.
- We have to use an existing CSDGM element that will be agnostic to the PID content.

- We will use a specific instance of Theme Keyword Thesaurus (USGS PID) and Theme Keyword for this purpose.
Where do I insert the PID in my CSDGM metadata record?

CSDGM does not provide a universal and unique field for a metadata identifier

- We have to leverage an existing element that can be leveraged for this purpose, without affecting validation rules.
- We will use an instance of thematic keyword/thematic keyword thesaurus to accomplish this:
  - Themekt=USGS Metadata Identifier
  - Themekey=USGS:uniqueIdentifierString
Where do I insert the PID in my ISO record?

- ISO 19115-X has a dedicated field for this very purpose:
  - Metadata Information :: MI_Metadata :: MI_Identifier
  - Because this field is reserved specifically for a metadata unique PID, you won’t need to insert the ‘USGS PID’ designation required for the CSDGM kluge.
When will this requirement be in place?

- PID Tool is expected to be complete within the next 60 days.
- OME integration should also be complete within the next 60 days.
- Once available, you should begin using it to assign a PID to your NEW metadata records.

- When SDC 3.0 launches at the end of FY20, it will scan all harvested records for the presence of the unique PID. Records that do not contain the PID will be rejected and not harvested into the SDC.
What about all of our already published metadata records?

- Fear not. You will not need to go back and assign PIDs to metadata published before the PID Tool becomes available (but you are welcome to do so!)

- The SDM Team will be taking on the task of retrospective assignment over the course of FY20.
  - ScienceBase data releases will be updated one by one.
  - WAF providers: we will automate updates to your WAF collections and push back to you new versions with PIDs.
  - IMU users: we will update the version in the IMU WAF with a PID, and push back a copy to the metadata contact via email.

- **Important note:** after these updates, the authoritative version of your metadata will be in ScienceBase or your WAF. You will need to download that version to perform future updates to your record, in order to preserve the persistence of the PID in the record.
What do we need from metadata reviewers and data managers?

- Help spread the word in your center/program about this forthcoming new requirement for USGS metadata.
- Once the PID Tool is released, begin using it for new metadata records.
- Make it part of your workflow to check for the presence of the PID in new metadata you are asked to review.
Questions?

Thanks for your time today!