2018 SWWG Meetings

October 11, 2018

Semantic approaches to enable USGS data to be FAIR (Findable, Accessible, Interoperable, Reusable)

We used the list of FAIR Principles at https://www.go-fair.org/fair-principles/, which includes links to explanations (links in the left column). Notes on the discussion are arranged by principle number:

F1. (Meta)data are assigned a globally unique and persistent identifier

Discussion was about separate metadata identifiers. In some cases the DOI for a data release might be considered the metadata DOI, or the IGSN for a sample might point to the metadata. Advantages of an identifier for the metadata are that it would be a way of discovering the most current version of the metadata, when the metadata is separated from the data landing page, and would be useful for managing collections of metadata.

F2. Data are described with rich metadata

Do USGS scientific communities have specific guidelines for "rich metadata"? The CSDGM Biological Profile is one. For data submitted to NWIS or Genbank, such guidelines are provided. In many cases, this guidance is provided by reviewers. The CDI project working on content specifications for ISO metadata is working on this.

F3. Metadata clearly and explicitly include the identifier of the data they describe

When data repositories assign identifiers, are they inserted in the metadata? The DOI could go into the online linkage.

F4. (Meta)data are registered or indexed in a searchable resource

The Science Data Catalog takes care of this!

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

Covered by the Public Access Policy and requirements for Trusted Digital Repositories.

A2. Metadata are accessible, even when the data are no longer available

A new USGS policy is needed. The metadata could continue to be provided on the landing page associated with the DOI.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

One model: NWISWeb provides RDF snippets in the XML metadata that include URIs for detailed descriptions.

I2. (Meta)data use vocabularies that follow FAIR principles

USGS vocabularies have a lot of features that are needed to follow FAIR principles. The problem is institutional commitment to a stable Web home for URI stability. We also need cross-vocabulary mapping especially for interdisciplinary science and conversations.

** This is an important topic for promoting data interoperability in USGS.

I3. (Meta)data include qualified references to other (meta)data

We need vocabularies and placement standards for specifying relationships with the items in references. It will be possible in ISO metadata. ScienceBase specifies it on landing pages, but that needs to be pulled into metadata for stand-alone use.

** This is an important topic. We could start by looking at DataCite specifications, the ScienceBase term list, and the ADIwg term list.

R1.1. (Meta)data are released with a clear and accessible data usage license

We have use constraints, but nothing called a license. We tend to assume people know they can use our data, but even saying “Public Domain” doesn’t specify what it means. CDI Software group is talking about it. Does ISO have a way? Does Lisa Zolly know more about what is possible?

** This is a thorny issue.

R1.2. (Meta)data are associated with detailed provenance

Provenance can be a huge amount of information. There is a Provenance Challenge in ESIP (linked below).

** This is an important topic to address.

Here are additional links from community members:

Link related to R1.2 Provenance: http://www.esipfed.org/community-prov-challenge

September 13, 2018

FY18 has been a quiet year for the Semantic Web Working Group. Do we still need the Semantic Web Working Group? If so, what do we need it to do?

A small group talked about the history and status of the working group, and concluded that we have achieved our original purpose of learning about semantic web technologies. We are ready to start assisting USGS in taking semantic approaches that will improve data management and integration. Improving USGS use of controlled vocabularies is an obvious place to start, and we agreed to encourage Peter Schweitzer to meet with us to identify opportunities.

For the FY19 CDI proposal cycle, we discussed proposing to host a USGS-wide workshop to design a road map to make USGS data more consistent with the FAIR Data Principles – not just focusing on integrating data to support a particular use, but improving our data practices so that all USGS data is findable, accessible, interoperable, and re-usable for multiple unanticipated uses.

Initial work on this proposal will happen over email.

June 14, 2018

We explored Loterre together.

Inist-CNRS (France) is rolling out Loterre (Linked open terminology resources), a multidisciplinary linked-data platform to make scientific terminology sources available on the web.

To facilitate exchanges and interoperability it is built on triplestore*and provides for resources retrieval and downloading together with a query API.

Loterre is not restricted to the terminology resources Inist-CNRS produces. It has also been designed as an open platform to host terminological data from other data producers.

The Inist also provides technological support for those who need to convert resources into SKOS/RDF format.

**A triplestore is a database especially designed for the storage and recovery of RDF data (Resource Description Framework)*

See Loterre: https://www.loterre.fr

Contact: contact-terminologietal@inist.fr

Stay tuned to Loterre on Twitter: https://twitter.com/INIST_Loterre

May 10, 2018

News sharing and discussion.

Notes on topics discussed:

- Fran and Matt saw a presentation that included an interesting use of ontologies and semantic tags, the use case being an international repository of ocean observing best practices. Fran will see if the presenter can speak to us about the technologies that are used, which might be useful in USGS for similar use cases. Some links: Dr. Pier Luigi Buttigieg is building the Environment Ontology, EnvO. On May 8, 2018, Buttigieg was part of a presentation about the Ocean Best Practices repository which is hosted by the UNESCO/IOC International Oceanographic Data and Information Exchange. The repository will use the EnvO ontology, to provide a "Semantic Advanced Search" capability that assists repository users in discovering best practices that meet their needs. A recording of the May 8 presentation is available online; a website summarizing the ocean observing best practices project is also available.
- The next CDI workshop will probably be June 2019, and there is some talk of having a controlled vocabulary "track" on one day of the workshop. We are interested in having a presentation, maybe a "birds of the feather" gathering, but it seems premature to separate out the controlled vocabulary people into a track. Instead we would like to participate actively in multiple programs for the purpose of raising awareness of the value of controlled vocabularies for multiple purposes.
- Might QMS kick-start our idea of a data dictionary database? QMS is the new Quality Management System that has been implemented for the energy laboratories, and will be implemented for other USGS laboratories over the next couple of years. Documentation of data-generating methods and review of data are part of the system. Data managers need to be active partners in implementing QMS and, in turn, the documentation of data dictionaries might serve as a starting place for compilation of a database that would be useful for USGS scientists and serve as a foundation for data integration.
- Ken and his colleagues are engaged in bridging the gap between disciplinary science and the computer science/linguistics/philosophical worlds of semantics. A journal article is planned, a blog post has already been released at http://aries.integratedmodelling.org/?p=1458. Ken welcomes our feedback about the blog post. We might also join him in writing something specific to USGS - e.g., "what is the semantic web, and how can it help us do cutting-edge science."
- Fran attended an webinar from the Association for Information Science and Technology (ASIS&T) about a new Dublin Core Metadata Initiative website that provides a competency framework for professional development in the use of linked data. The site also provides links to selected educational materials and a dataset for use in assigning or completing exercises. It is freely available at explore.dublincore.net. We might want to explore it together as part of this summer's CDI training program. Fran pointed out that the "Creating linked data applications" topical cluster in the competency framework is very undeveloped, which matches our experience that the major problem in semantic web implementation is providing applications that exploit the potential power of semantic web technologies.

April 12, 2018

Demonstrating the use of permanent identifiers in linked data

Example:

Data Categories for Marine Planning (DCMP), a vocabulary of 90 terms (https://pubs.er.usgs.gov/publication/ofr20151046)
Triple store of DCMP terms (created by Rensselaer Polytechnic Institute): https://nocv.tw.rpi.edu/elda/api/vocab/dcmp/terms

In the human-readable interface of the triple store, each of the 90 DCMP terms has its own URL:

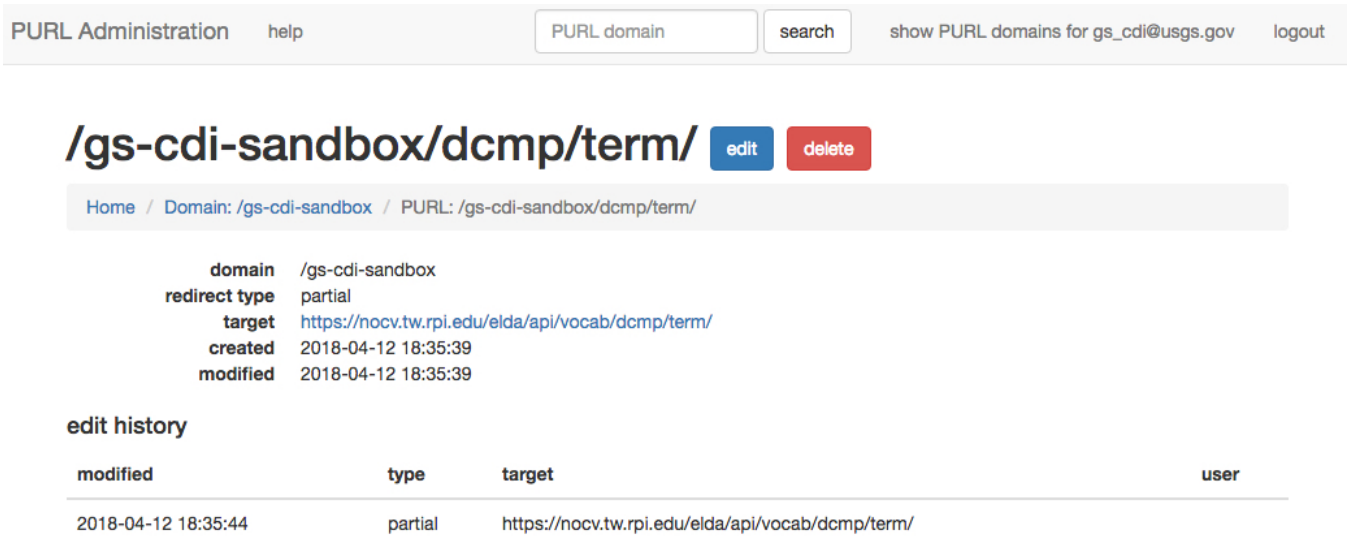
https://nocv.tw.rpi.edu/elda/api/vocab/dcmp/term/assessments
https://nocv.tw.rpi.edu/elda/api/vocab/dcmp/term/bathymetry-and-elevation
https://nocv.tw.rpi.edu/elda/api/vocab/dcmp/term/biodiversity
https://nocv.tw.rpi.edu/elda/api/vocab/dcmp/term/biological-occurrence
https://nocv.tw.rpi.edu/elda/api/vocab/dcmp/term/biological-production
.
.
.
etc.

... but the linked data files themselves (RDF/XML, N3, JSON, and other formats) use w3id permanent identifiers (https://w3id.org/) that point to those URLs:

https://w3id.org/national-ocean-council/api/vocab/dcmp/term/assessments
https://w3id.org/national-ocean-council/api/vocab/dcmp/term/bathymetry-and-elevation
https://w3id.org/national-ocean-council/api/vocab/dcmp/term/biodiversity
https://w3id.org/national-ocean-council/api/vocab/dcmp/term/biological-occurrence
https://w3id.org/national-ocean-council/api/vocab/dcmp/term/biological-production
.
.
.
etc.

DEMONSTRATION: Creating a new set of permanent identifiers using the PURL system (https://archive.org/services/purl/). Leslie Hsu kindly offered the CDI PURL "sandbox" account for the demo.

In the demo, Alan created one "partial" PURL: http://purl.org/gs-cdi-sandbox/dcmp/term/ that redirects to the target URLs beginning with: https://nocv.tw.rpi.edu/elda/api/vocab/dcmp/term/ (see screenshot, below)



... and the PURL resolver redirects to each of the 90 individual terms:

http://purl.org/gs-cdi-sandbox/dcmp/term/assessments
http://purl.org/gs-cdi-sandbox/dcmp/term/bathymetry-and-elevation
http://purl.org/gs-cdi-sandbox/dcmp/term/biodiversity
http://purl.org/gs-cdi-sandbox/dcmp/term/biological-occurrence
http://purl.org/gs-cdi-sandbox/dcmp/term/biological-production
.
.
.
etc.

Why bother? In this example, we might want to export the DCMP linked data from the RPI triple store and import it into a USGS triple store. If we do that we'd need to replace the RPI-controlled permanent identifiers with USGS-controlled permanent identifiers.

Another example of why it's advisable to use permanent identifiers in linked data: Alan and Fran are co-authors of an OFR that provides a shapefile of the spatial extents (as simple polygons) of 300 undersea features (https://pubs.er.usgs.gov/publication/ofr20141040). Version 1.1 of the OFR, to be published soon, will also include a linked-data representation of the polygons (in WKT). For now, the PURLs in the linked data will point to the OFR, but if USGS establishes a permanent triple store for linked data, we may want to change the targets of the PURLs without having to revise the linked data files themselves.

We may continue this discussion at a later SWWG meeting.

January 11, 2018

Planned agenda:

1. Experiment with Google Hangouts
2. Continuing last month's discussions.

Attending: Alan Allwardt, Peter Schweitzer, Ken Bagstad, Fran Lightsom, Andy LaMotte, Dave Govoni, Matt Arsenault, Mike Ierardi, Dave Coyle

CDI proposal discussion: What would be the benefit of doing the CDI project that would make it worthwhile to jump through the hoops? What would we do with the CDI money and where would we get the matching funds? Some of us need funding for our time to participate in a project. The rest of us could provide our time as matching.

- Data dictionary database: take a year to experiment and flounder around and maybe propose next year.
- Semantic meta-modeling: Ken would lead development of SOI.

Getting organized for respondent panel at Feb. CDI meeting. Ken will get preview of presentation. Fran will recruit some more panel members and invite the panel to a preliminary discussion on about Feb. 1.

Google Hangout worked pretty well, after some fumbling around at the beginning of the call.