Community Mediation Service for Provenance and Annotation **[1]**

Submitted by sbristol on Sun, 2016-08-14 09:32

Overview

The science conducted by Federal agencies is becoming increasingly interdisciplinary and synthesis focused as we work to answer questions of broad scope and major societal importance. From a data and information perspective, scientific synthesis involves the combination and integration of assets from across many agencies, universities, and other contributors. Being able to trace back to the source of the material used in scientific findings is fundamental to transparency, traceability, and reproducibility of the scientific endeavor along with ensuring quality, objectivity, utility and integrity of the scientific process - foundations of the 2001 Information Quality Act (IQA).

Data and information management efforts such as the Global Change Information System and recent efforts in the USGS to produce a Biogeographic Information System are often working retrospectively to curate provenance information and capture annotation about scientific data and information assets. We do this work in order to produce assessments and other synthesis products that provide trustworthy, actionable science to policy makers and the public. Steadily, we are beginning to develop data and information networks coupled to advanced technologies that are capable of capturing and recording usable logs for what is happening in the production and analysis of data, the publishing of scientific results, and the steady growth of scientific knowledge. These contributions should encourage future efforts to begin looking to how we can bring together provenance information and meaningful annotation in near real time with curation activities focused more on methods of summarizing information to different levels or contextual forms to meet particular user needs.

Perspective

In August 2016, the USGS is launching a new ESIP Testbed activity to explore a community-based provenance and annotation mechanism capable of combining information where more than one repository is recording useful information about data products and processing; relevant information for cross-community use. For instance, one of our use cases involves the processing of marine biological observation data collected by organizations funded through the NOAA Integrated Ocean Observing System (IOOS). IOOS Regional Associations operate Data Assembly Centers (DACs) and provide initial processing of the data to align them with the Darwin Core standard for biological data, using an implementation that IOOS calls the Biological Data Services. Original source data can be something as simple as a set of human observations obtained using a protocol like the visual line transect survey to something more complex such as passive acoustic sonar data or eDNA that require significant processing to identify species or taxonomic group. Processed datasets are served via ERDDAP from the IOOS RAs, and then retrieved by the USGS-operated Ocean Biogeographic Information System (OBIS) for USA to perform final quality control checks and further processing to align taxonomy with the World Register of Marine Species (WoRMS) and/or the Integrated Taxonomic Information System (ITIS) to add a scientific name identifier to the data. The data are then served up to the International OBIS and the Global Biodiversity Information Facility (GBIF) for broader access and use. Accessed through OBIS and GBIF at the terminal points of data evolution and processing, users are often picking up an aggregate dataset from across many different sources, and they ultimately need a record-level provenance trace that helps them understand where the data originated and what all happened to the data along the journey.

As data retrievals reach in the millions of records needed for a particular type of analysis, and as data stores across multiple domains expand with exponential growth, it has become imperative that the provenance trace be structured in such a way that the user can integrate this information, guided in tandem with other relevant criteria, to tease out or facet in on a particular subset that is most appropriate for their particular use.

Concept

Capturing downstream information in the data life cycle, in cases such as OBIS and GBIF, can be encoded using the W3C-PROV standard that has seen much attention in the ESIP community. PROV statements are often most powerful when entities, agents, and even actions are recorded as persistent identifiers from some other actionable source such as an ontology, registry, or catalog. That “shorthand” for the various parts of a provenance trace can give us powerful ways of querying and using the data, including methods for assembling human readable provenance notation when it’s needed. For the ESIP Testbed we are working on, we envision close connection with the ESIP Community Ontology Repository along with other activities involving persistent identification of resources of various types.

One of the other aspects of our overall challenge of multiple organizations operating on the same data and information in various ways may be described as annotation. There are cases where an annotation concept is directly part of a data schema and can simply be recorded within the data as things move along. However, there are many cases where annotation is something related to but apart from data, information, scientific software, or some other artifact in the overall research infrastructure. A provenance statement may record that a particular person or an algorithm of some kind made an annotation about the entity it was operating on. It may be possible to store the content of that annotation within the PROV structure, but it may or may not be the most viable and usable way for recording that information. The W3C candidate recommendation for annotation is a very simple model that essentially sets up a triple, connecting annotation content with the subject/target of the annotation with a particular annotation type designation.

Annotation, like provenance, is an information resource that is created as part of the overall scientific workflow. Both annotation and provenance are produced by different people and processes across organizations, but they are often produced by the same agents, about the same entities, using the same actions, and on the same targets. In these cases, we need an ability to trace back through and assemble everything we know about that workflow. To the extent that PROV and annotation are based on persistently identified concepts in the ontology space, perhaps through ontologies registered in the ESIP COR, we should be able to follow the links and create one or more usable representations of the data for various purposes. That capability is the subject of our exploration under the USGS Testbed activity with ESIP. We have several specific use cases we will be exploring, including the OBIS case described above, and we will be looking for some other interesting cases to emerge through discussions in the ESIP community.

The architecture used to establish a community mediation service or capability could follow a number of different patterns from being highly centralized to widely distributed. We are leaning more toward the latter based on experience and practicality where we would end up with more of a pull from registered systems vs. everyone pushing their information to some central place. Provenance and annotation information will ultimately be stored in very large pools of resources. Each organization from data repositories to analytical labs that might participate in such an enterprise will have its own operational reality to deal with. There will be a variety of technologies in use just like there are in any other areas we are trying to connect such as with metadata catalogs. We are generally thinking, at this point, that multiple organizations will generate and store provenance (PROV) and annotation in various ways but expose compatible APIs that follow a constrained set of standards and/or conventions using the W3C models. Those APIs will need to advertise themselves or be registered in some way, which will be one of the key aspects of the community mediation concept. Looking across sources could involve anything from a distributed search to some form of federation (virtualization) or indexing. Any given “member” of the notional network might opt to run its own federation or select those of other members based on local needs.

Next Steps

To pursue the architectural questions and other interesting issues like various ways that provenance

might be summarized, synthesized, or distilled into tractable forms when it moves out of a given source, we plan to assemble a team of experts across multiple disciplines and with different viewpoints into a synthesis working group. Building off this basic “idea document,” we will be laying some groundwork for well described use cases prior to the January 2017 Winter Meeting. Sometime shortly into the new calendar year, we plan on holding a focused working group session where we invite folks in a room for a few days to work through some questions and set up working software. We’ll articulate some experiments based on our use cases during that session and then proceed with testing extending over a couple of months. Following that, we’ll meet again, review the results, write up what we learned, and make some recommendations on where our community-based provenance and annotation capabilities might go from there.

Short name: usgs_provannotation
Project type: Full project
Enable issue tracker: No

Source URL: https://testbed.esipfed.org/node/9350

Links
[1] https://testbed.esipfed.org/node/9350