Community-based cyber infrastructure research and development assistance
Provenance Activity Report
June, 2018

Problem Statement

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 the source of all materials 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 Information Quality Act (IQA)[1].

Data and information management efforts such as the Global Change Information System[2] and recent efforts in the USGS to produce a Biogeographic Information System are often working retrospectively to curate provenance[3] information and capture annotation about scientific data and information assets.

Capturing downstream information in the data life cycle can be encoded using the W3C-PROV standard. 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.

Annotation is another way to describe the overall challenge of multiple organizations operating on the same data and information in various ways. 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.

[1] https://fas.org/sgp/crs/RL32532.pdf

[2] https://data.globalchange.gov/

[3] https://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance#A_Working_Definition_of_Provenance

Proposed Activity Review

To address these challenges, USGS and ESIP partnered on the "Community-based cyber infrastructure research and development assistance" award (Funding opportunity number: G17AS00001); a key component of which centered on data provenance. Specifically, the award stated that ESIP would: *partner with the USGS to evaluate an approach to two major aspects of the scientific workflow: 1) registering/recording the provenance of integrated and synthesized scientific data and scientific findings and 2) registering/recording annotations on scientific data and information assets made by the scientific community and by other interested parties.*

ESIP accomplished this goal through an event called the Community PROV Challenge: a three-part series of activities that moved through the process of idea gathering, prototype development, and working group formation.

Step One: Idea Gathering

The first stage of the Community PROV focused on ideation. Individuals contributed ideas to the IdeaScale platform answering the question, "How would YOU improve how W3C-PROV systems interoperate across agencies or institutions to enable a more complete picture of provenance?" Ideas were then voted and commented on, creating further dialogue around the challenge question. A follow-on breakout session was convened at the ESIP Summer Meeting, June 2017 in Bloomington, IN. The author of the idea that received the most 'up votes' received paid travel to the 2018 ESIP Winter Meeting.

Outcomes:

- 17 Ideas Posted
- 27 Comments
- 35 Votes
- 41 Individual Participants

Most popular idea was Visualization of Provenance Traces, submitted by Tom Narock: *Some simple visualization tools that will graphically show the lifecycle of a dataset would be very helpful. I'd suggest a web-based visualization service (perhaps using D3) that can aggregate related PROV and visualize the resulting data lifecycle. The service would dynamically generate a list of all datasets it knows about, users would select one, and the service would visualize the provenance. Additional features might include the ability to graphically compare two or more provenance traces and highlight differences.*

How exactly this is implemented is going to depend on the underlying inter-agency architecture and structure of the provenance documents.

Stage Two: Prototype Development

The ESIP Lab sought proposals from qualified teams to develop, extend, or fully test a prototype community-mediation capability for provenance and annotation generated and exposed by disparate sources but summarized, synthesized, or distilled into tractable forms for community use. Prototypes were based on a distributed system where organizations, from data centers to analytical labs, have

different means of and underlying technologies for generating and storing PROV and annotation, but expose compatible APIs that follow a constrained set of standards and/or conventions using the W3C specifications. The ESIP Lab awarded one project $15,000 to create a prototype solutions that was presented at the 2018 ESIP Winter Meeting.

Outcomes:

- Provisium.io
- Github Repo
- 2018 Winter Meeting Presentation (Narock)
- 2018 Winter Meeting Presentation (Fils)

Stage Three: Working Group Formation

To culminate the Community PROV Challenge and create a bridge to 'what's next' in the world of improved provenance and annotation at an interagency level, ESIP organized the Earth Data Provenance Workshop. The workshop was March 27 - 29, 2018 at the eScience Institute in Seattle, WA. This was a synthesis working group style activity focused on creative thinking around provenance.

The workshop agenda focused on understanding (and finding solutions to overcoming) the multiple impedance points that exist along the path to a fully-distributed provenance system for Earth science data. This workshop examined the provenance workflow from a distributed and generic perspective, with the intention of empowering workshop participants to move towards practical implementations of provenance.

The workshop convened technologists and academic researchers from across agencies, institutions and the data lifecycle. Combined, the participants represented a breadth of interest in provenance and annotation, which resulted in engaging, productive workshop deliberations.

The intention of this workshop was to act as a kickoff for two years of forward motion in the field of Earth data provenance. Beyond the Seattle workshop, participants will come together at the bi-annual ESIP meetings to discuss the state of Earth data provenance, showcase progress and create future action items.

Outcomes:

- Earth Data Provenance Workshop Report
- Earth Data Provenance Github Repo

Summary

The unique power and flexibility of the ESIP community has enabled the USGS to engage multiple partners from across the spectrum of Earth system science—government, academia, commercial and non-profit—around Earth data provenance. The Community PROV Challenge provided insights into emerging methodologies and technologies to manage information flows from multiple sources simultaneously and through time. Under a paradigm where Earth science researchers combine data from multiple agencies into new syntheses that answer questions, and drive further inquiry, far beyond what any one agency could pursue on their own, continued efforts around data provenance and annotation will be critical.