Formal Provenance Representation of the Data and Information Supporting the National Climate Assessment

Curt Tilmes

National Aeronautics and Space Administration
Curt.Tilmes@nasa.gov

Abstract. The Global Change Information System (GCIS) provides a framework for the formal representation of structured metadata about data and information about global change. The pilot deployment of the system supports the National Climate Assessment (NCA), a major report of the U.S. Global Change Research Program (USGCRP). A consumer of that report can use the system to browse and explore that supporting information.

Additionally, capturing that information into a structured data model and presenting it in standard formats through well defined open interfaces, including query interfaces suitable for data mining and linking with other databases, the information becomes valuable for other analytic uses as well.

1 Background

1.1 USGCRP and the NCA

The United States Global Change Research Program (USGCRP) was established by Presidential initiative in 1989 and mandated by Congress in the Global Change Research Act (GCRA) of 1990 [3] in order to "assist the nation and the world to understand, assess, predict, and respond to human-induced and natural processes of global change." That act also requires the program to submit an assessment (not less frequently than every 4 years) which:

1. Integrates, evaluates, and interprets the findings of the Program and discusses the scientific uncertainties associated with such findings
2. Analyzes the effects of global change on the natural environment, agriculture, energy production and use, land and water resources, transportation, human health and welfare, human social systems, and biological diversity
3. Analyzes current trends in global change, both human-induced and natural, and projects major trends for the subsequent 25 to 100 years.

There have been three National Climate Assessment reports, with the latest being released in 2014, "Climate Change Impacts in the United States: The Third National Climate Assessment."[2]

1.2 Provenance

The World Wide Web Consortium (W3C) Provenance Working Group defines provenance as "information about entities, activities and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness."

A goal for support of the NCA is to be capable of storing sufficient information about the findings and information in the NCA that a reader can explore that provenance and understand where they came from. This includes referenced scientific articles, datasets, models, observations and measurements and also the people, organizations, agencies, instruments, satellites, in-situ sensor networks, etc. that contributed to that information.

1.3 Global Change Information System

The Global Change Information System (GCIS) is a web-based resource for traceable, sound global change data, information, and products. Designed for use by scientists, decision makers, and the public, the GCIS provides coordinated links to a select group of information products produced, maintained, and disseminated by government agencies and organizations. As well as guiding users to global change research products selected by the 13 member agencies, the GCIS serves as a key access point to assessments, reports, and tools produced by the U.S. Global Change Research Program.

The first instantiation of the GCIS presents the Third NCA report, and includes a structured representation of information and data supporting the findings and figures of that report. Though much of the focus of the GCIS is on the "front-end" presentation of the report and the supporting information, this paper will describe the "back-end" archive and interfaces to that structured information in the pilot deployment of the system at http://data.globalchange.gov.

2 GCIS Implementation

2.1 Resources and Identifiers

Each item or resource in the GCIS has a unique, persistent, resolvable identifier. We want to follow the linked data principles as described[1] by Tim Berners-Lee, so we use URIs (actually, "Cool URIs"[1]) as identifiers. We also want to control the resulting information when the identifier is resolved, so we have rooted our identifiers in our own http://data.globalchange.gov namespace. We'll call those the "Global Change IDentifiers" or GCIDs.

Many of the items are "owned" by other entities and have perfectly good universal identifiers assigned in various systems. For example, many of our resources are journal research articles. We consider them "owned"[2] by the publisher of the

[1] http://www.w3.org/TR/cooluris/

[2] For the purposes of this discussion, we intentionally ignore the complexities of intellectual property rights.

article. The publisher has typically assigned a Digital Object Identifier (DOI) to the article that when resolved through the DOI resolver, http://dx.doi.org forwards the requester to the publisher's landing page for that particular article.

An example of this is this GCID

```
http://data.globalchange.gov/article/10.1038/425365a
```

which refers to a particular article *Oceanography: Anthropogenic carbon and ocean pH* by Caldeira and Wickett, assigned `doi:10.1038/425365a` by the publisher, which when resolved with

```
http://dx.doi.org/10.1038/425365a
```

forwards to more information about the article here:

```
http://www.nature.com/nature/journal/v425/n6956/full/425365a.html.
```

Each resource in the GCIS is assigned a GCID. Some of the resources already represented in the system are: Report, Chapter, Figure, Image Finding, Table, Reference, Publication, Contributer (Organization or Person), Dataset and Activity. As our use case exploration continues more and more related resources are being added to the data model.

2.2 Relational Data Model

The heart of the GCIS is a relational data model that captures both basic information (metadata) about each resource, but more importantly, the relationships between resources. In particular, this includes provenance relationships such as a Report has a Chapter that has a Figure that has an Image that was generated by an Activity and derived from a Dataset and so on.

2.3 RESTful Interface

As described above, each resource in the GCIS has a resolvable URI (the "GCID"). When requested via HTTP, the URI will, by default, return a human readable expression of the metadata and relationships of that resource in HTML. These include links to other resources (via their GCIDs), or to API elements that list various types of resources related to that resource. For example, resolving the report GCID http://data.globalchange.gov/report/nca3 includes a reference to http://data.globalchange.gov/report/nca3/chapter, which lists the individual chapter resources that are part of that report.

Given proper authentication credentials and authorizations, the interface also supports HTTP `PUT`, `POST`, and `DELETE` methods to round out the RESTful CRUD (Create, Read, Update and Delete) operations.

The interface also supports other more structured formats for the information returned for each GCID. These include JSON (easily parsed and used by Web front ends to depict the information), YAML (easy to read by both humans and machines), and several expressions of Resource Data Format (RDF) triples.

2.4 SPARQL Endpoint

Finally, all of the RDF triples derived from all of the structured information from all of the GCIS resources are exported into a triple store queryable through a public SPARQL interface http://data.globalchange.gov/sparql. While the web interface to the structured information supports basic "browsing" through the provenance (from the Report page, go to a Chapter page, to a Figure page, to a Dataset page, to the data center distributing that data), the SPARQL interface supports more interesting data mining type queries such as "Which findings of the NCA report are supported by datasets derived from observations that came from the Aqua satellite launched by NASA?" or looking in the other direction "Which research articles support the NCA finding that extreme weather events are increasing in frequency?"

It could also provide metrics like "How many research articles cite data derived from observations from each of NASA's satellite missions?"

3 Discussion

In its pilot form deployed to support a single report, the GCIS is still very incomplete, both in its model and its content. The basic framework as deployed does demonstrate, however, the potential of this approach to capture and present some incredibly useful information.

Tiny bits of the provenance information about reports such as the NCA are typically strewn across a wide swath of the scientific literature, agency web sites, natural language expressions in articles and reports, with a great deal of it existing only on personal computers and even in researchers heads.

Providing a framework like GCIS and a data model for capturing, expressing, and analyzing those holdings will help us begin to extract that information, transform it into machine readable, semantic form and consolidate it into a mineable database supporting a wide variety of analyses.

In addition to completing the huge task of documenting the provenance and supporting information for the Third National Climate Assessment, we plan to extend the GCIS to other upcoming reports and also extend the data model to support other use cases. Ultimately it should be possible to connect the universe of information about global change from observations to data to research to assessments and findings so researchers can understand how it all fits together.

References

1. Berners-Lee, T.: Linked data (2006), http://www.w3.org/DesignIssues/LinkedData.html
2. Melillo, J.M., Richmond, T.T., Yohe, G.W. (eds.): Climate Change Impacts in the United States: The Third National Climate Assessment. U.S. Global Change Research Program (2014), doi:10.7930/J0Z31WJ2
3. U.S. Code: Global Change Research Act of 1990 (P.L. 101-606) (1990)