Supporting Trust with Provenance of the Findings of the National Climate Assessment

Curt Tilmes[1], Robert E. Wolfe[1,2], Brian Duggan[2,3], Steven Aulenbach[2,3], Justin C. Goldstein[2,3], Xiaogang Ma[4], and Stephan Zednik[4]

[1] NASA Goddard Space Flight Center, Greenbelt MD, USA
{Curt.Tilmes,Robert.E.Wolfe}@nasa.gov
[2] U.S. Global Change Research Program, Washington DC, USA
{rewolfe,bduggan,saulenbach,jgoldstein}@usgcrp.gov
[3] University Corporation for Atmospheric Research**, Boulder CO, USA
[4] Rensselaer Polytechnic Institute, Troy NY, USA
{max7,zednis2}@rpi.edu

Abstract. A typical scientific assessment report such as the USGCRP National Climate Assessment [5] considers the corpus of research in the field and through a deliberative process distills and synthesizes that knowledge into specific findings documented and supported in the report. Human readers of the report can determine trust by examining supporting and related aspects of the report such as the producing organization, the credentials and expertise of the authors, the development and review process, referenced scientific research, etc.
This paper describes an approach for encoding some of those provenance artifacts supporting the findings in machine readable structured forms. This encourages a very formal and rigorous process for the support of those findings, and enables more systematic exploration of the provenance of the findings of the report. We also share a public open dataset with such information supporting the findings of the National Climate Assessment in the Global Change Information System.

Keywords: Trust · Provenance · Assessment

1 Introduction

The U.S. 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 to "assist the Nation and the world to understand, assess, predict, and respond to human-induced and natural processes of global change." The USGCRP is required by U.S. law to submit an assessment of the overall findings of the program. This paper will focus on several key aspects of the report development process [1] that enable documentation of provenance of

** This material was developed with federal support through the U.S. Global Change Research Program under National Science Foundation contract no. NSF-DACS13C1421.

the findings of the USGCRP National Climate Assessment (NCA) in innovative ways supported by a **Global Change Information System** (GCIS) and a machine readable representation of that provenance that could support trust exploration and analysis and quality assessment of the findings of the NCA.

The most recent NCA, *Climate Change Impacts in the United States: The Third National Climate Assessment* [5] can be read online at `https://nca2014.globalchange.gov` and the supporting provenance can be explored at `https://data.globalchange.gov/report/nca3`.

2 National Climate Assessment

The NCA report itself comprises numerous chapters, figures, findings, tables and references. The main chapters each had individual lead authors and contributing authors who performed the work of the assessment and developed their chapters. Contained within the chapters are specific findings that are supported by figures tables and references.

For each of the findings in the report, the authors documented the process and details about the development of the findings in a "traceable account" of that finding. These include narrative descriptions supplied by the authors:

- Process for developing key messages
- Description of evidence base
- New information and remaining uncertainties
- Assessment of confidence based on evidence

For example, one of the findings from chapter 2 "Our Changing Climate" is:

> *U.S. average temperature has increased by 1.3°F to 1.9°F since record keeping began in 1895; most of this increase has occurred since about 1970. The most recent decade was the nations warmest on record. Temperatures in the United States are expected to continue to rise. Because human-induced warming is superimposed on a naturally varying climate, the temperature rise has not been, and will not be, uniform or smooth across the country or over time.*

The complete traceable account describing the process and evidence for that finding and justifying the ***very high*** confidence in that finding can be reviewed at `https://data.globalchange.gov/report/nca3/chapter/our-changing-climate/finding/us-temperature-increased`.

3 System Description

The GCIS maintains two concurrent models of the data: a relational model and a semantic model. (`http://data.globalchange.gov/resources`) These each describe the resources of the system (e.g. Report, Chapter, Figure, Finding, Person, etc.) and their interrelationships. Every resource is assigned a GCIS

```
<http://data.globalchange.gov/report/nca3>
   dcterms:identifier "nca3";
   dcterms:title "Climate Change Impacts in the United States: The Third
      National Climate Assessment"^^xsd:string;
   gcis:hasChapter <http://data.globalchange.gov/report/nca3
      /chapter/our-changing-climate>;
```

Fig. 1. http://data.globalchange.gov/report/nca3

```
<http://data.globalchange.gov/report/nca3/chapter/our-changing-climate>
   dcterms:identifier "our-changing-climate";
   gcis:chapterNumber "2"^^xsd:integer;
   dcterms:title "Our Changing Climate"^^xsd:string;
   gcis:isChapterOf <http://data.globalchange.gov/report/nca3>;
   gcis:hasFigure <http://data.globalchange.gov/report/nca3/chapter
      /our-changing-climate/figure/observed-us-temperature-change>;
   gcis:hasFinding <http://data.globalchange.gov/report/nca3/chapter
      /our-changing-climate/finding/us-temperature-increased>;
   prov:qualifiedAttribution [
      a prov:Attribution;
      prov:agent <http://data.globalchange.gov/person/1009>;
      prov:hadRole <http://data.globalchange.gov/role_type
         /convening_lead_author>;
      prov:actedOnBehalfOf <http://data.globalchange.gov/organization
         /university-alaska-fairbanks>;
      ] ;
```

Fig. 2. http://data.globalchange.gov/report/nca3/chapter/our-changing-climate

```
<http://data.globalchange.gov/report/nca3/chapter/our-changing-climate
      /finding/us-temperature-increased>
   dcterms:identifier "us-temperature-increased";
   gcis:findingNumber "2.3"^^xsd:string;
   dcterms:description "U.S. average temperature ..."^^xsd:string;
   gcis:isFindingOf <http://data.globalchange.gov/report/nca3
      /chapter/our-changing-climate>;
   gcis:isFindingOf <http://data.globalchange.gov/report/nca3>;
   gcis:findingProcess "Development of the key messages ..."^^xsd:string;
   gcis:descriptionOfEvidenceBase "The key message and ..."^^xsd:string;
   gcis:assessmentOfConfidenceBasedOnEvidence
      "Given the evidence base and remaining uncertainties..."^^xsd:string;
   gcies:newInformationAndRemainingUncertainties
      "Since the previous National Climate Assessment, there..."^^xsd:string;
   cito:cites <http://data.globalchange.gov/article/10.1029/2011JD016761>;
   biro:references <http://data.globalchange.gov/reference/
      66ccff5f-4828-4e03-be08-ee6f49296f34>;
```

Fig. 3. .../chapter/our-changing-climate/finding/us-temperature-increased

identifier (GCID) which is a resolvable URI for that resource. A simple RESTful API (`https://data.globalchange.gov/api_reference`) allows interaction with those resources. Public HTTP `GET` on the resource will retrieve it, and given proper credentials, HTTP `PUT` can be used to create or modify those resources.

Following a linked data approach, each resource has been assigned a specific, persistent, resolvable, referenceable identifier, referred to as a Global Change Identifier or GCID.[2] Typically when the components are nested within a higher level component (e.g. chapters of a report), the GCIDs also nest their URIs. When the components are externally relevant (e.g. a person or organization), they are not nested. Some example GCIDs related to the NCA are shown in Table 1.

Table 1. Example GCIDs

`https://data.globalchange.gov/report/nca3`
`https://data.globalchange.gov/report/nca3/chapter/our-changing-climate`
`https://data.globalchange.gov/person/1009`
`https://data.globalchange.gov/organization/university-alaska-fairbanks`
`.../report/nca3/chapter/our-changing-climate/figure/observed-us-temperature-change`
`.../report/nca3/chapter/our-changing-climate/finding/us-temperature-increased`
`.../report/nca3/reference/32bec5d2-97fe-41c5-8eed-6920bbf096f4`
`.../article/10.1175/2008JCLI2263.1`

Various metadata about each resource are captured and represented in the system.[3] That information is presented in both an HTML format which can be rendered in a user friendly form, and also in various machine readable formats. These include JSON which makes it very easy to build web clients for displaying and interacting with the system through the API, and RDF based formats suitable for linking with external items in the linked data or semantic web domains.

The system allows both URL suffix and HTTP Accept header to request content negotiation and supply the various formats. For example, a JSON description of the structured elements of the report can be retrieved from `https://data.globalchange.gov/report/nca3.json`, and a turtle description from `https://data.globalchange.gov/report/nca3.ttl`. The turtle format data is produced via handcrafted templates populated from the appropriate fields of the RDBMS. The other RDF formats are created via on-the-fly conversion from the turtle. As an added benefit for machine exploration of the data, the RDF data are periodically exported by walking the entire database and imported into a triple store with public SPARQL access using the GCIS ontology. [4] (See: `https://data.globalchange.gov/sparql`)

Some brief, truncated examples of the turtle formats are shown in Figures 1, 2 and 3 (see the actual URIs above for more complete examples). For the finding in figure 3, the fields from the traceable accounts are included, as well as structured supporting information are linked to the GCID. For this initial effort, the full text of narrative fields from the traceable accounts of the findings were included verbatim as literal strings in the RDF. Future efforts may try to

tease these out in a more structured form. Each of the supporting references are linked as shown as well.

For the NCA report and some of the other projects now using the GCIS framework, we've used a variety of mechanisms to add resources and relationships to the system. Where external databases exist with already structured information, external "synchronization modules" (`https://github.com/USGCRP/gcis-sync`) are used to query those databases and exercise the GCIS API to add/update information. These run periodically to harvest information maintained by trusted partners. Sometimes information has been captured by staff in various spreadsheets or other structured forms and a hodgepodge set of scripts transform those data and interact with the GCIS API as well.(`https://github.com/USGCRP/gcis-scripts`) Finally various client-side forms can be used to manually enter information into the system given proper credentials.

Regardless of the mechanism used to ingest information into the GCIS through its API, a complete audit trail of the credentials used to enter the information is logged. There are also multiple instances of the GCIS system, with the operational public instance at `https://data.globalchange.gov` being read/only, with changes going into a "stage" instance where a curation process promotes that information into the operational instance.

4 Lessons Learned

The author team (over 250 directly acknowledged authors, but many more contributing to the technical inputs) for the report was quite diverse and it was very challenging to bring consistency to the process of development. As very senior publishing scientists, they are very familiar with traditional publishing paradigms, but less so with the highly structured approach described here. For example, when referencing another journal article, if the citation is intended solely for a human audience, and particularly scientists very familiar with the literature of the field, the detailed metadata of the cited article needn't be incredibly precise. If, however, we want a computer to be able to parse the metadata, understand the particular referenced article, and even retrieve additional metadata from the publisher's site, we need very precise metadata. With the pervasive use of structured metadata, automated reference managers and digital object identifiers (DOIs) in the scientific publishing industry, one might assume this problem has long been solved, but even today the quality of references provided by authors was not sufficient for our needs and required extensive automated and manual quality checking and many rounds of revision. Documenting other types of provenance, and deeper provenance for NCA3 has been an extremely challenging, very manual process.

Now that the basic framework is in place, however, it has been easier to demonstrate the potential of this system, and convince contributors of the usefulness and describe what will be needed prior to their development of the report. In addition to plans to use this system to support the next major National Climate Assessment report, there are several other reports in production which will

use it as well. In particular, a report in development, *The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment* will make use of this system and those authors are already being trained.

To address some of the problems described above, the team is trying to engage the author teams earlier in the process and use a variety of training and tools to inform the teams of the needs, and make it easier for them to supply compliant information in forms more easily incorporated into the system. Improving that training and developing and enhancing those tools will hopefully improve the whole process resulting in a higher quality product for the future.

5 Conclusions

Each of the provenance artifacts supporting the findings of this assessment report have been assigned resolvable, linkable URI identifiers (GCIDs), including 43 chapters (chapter-like resources, citing 3,354 references), 290 figures (comprising 458 individual images, citing 232 references), 161 findings (citing 1,195 references), 20 tables (citing 104 references) and 3,395 total references. Many of those artifacts are further linked to their supporting provenance (e.g. a figure with a graph or map is linked to the dataset from which it is derived).

This approach allows each of those artifacts to be linked and referred to, and the HTML based user interface allows the typical "follow your nose" exploration of the structured information about the report and drilling down into the supporting information. It also provides a real world example of encoding that provenance into a structured, machine readable format (utilizing the GCIS [4] and PROV [6] and other ontologies) that can be used for queries and data mining and exploration of additional concepts for analyzing trust based techniques in a real world use case.

References

1. Buizer, J.L., Fleming, P., Hays, S.L., Dow, K., Field, C.B., Gustafson, D., Luers, A., Moss, R.H.: Report on Preparing the Nation for Change: Building a Sustained National Climate Assessment Process (2013)
2. Duggan, B., Tilmes, C., Aulenbach, S., Wolfe, R.E., Goldstein, J.C., Manipon, G.: Normalizing Resource Identifiers using Lexicons in the Global Change Information System. Proc. of Linked Data on the Web 2015 (2015)
3. Ma, X., Fox, P., Tilmes, C., Jacobs, K., Waple, A.: Capturing Provenance of Global Change Information. Nature Clim. Change 4(6), 409–413 (06 2014), http://dx.doi.org/10.1038/nclimate2141
4. Ma, X., Tilmes, C., Fox, P.: GCIS ontology version 1.2 (2013), http://dx.doi.org/10.7930/J0HQ3WTK
5. Melillo, J.M., Richmond, T.C., Yohe, G.W.: Climate Change Impacts in the United States: The Third National Climate Assessment (2014), http://dx.doi.org/10.7930/J0Z31WJ2
6. Moreau, L., Missier, P.: Prov-dm: The prov data model. Tech. rep. (2012), http://www.w3.org/TR/prov-dm/