PROV in the Global Change Information System (GCIS)

Curt Tilmes, Robert E. Wolfe

NASA Goddard Space Flight Center
Greenbelt, MD
Curt.Tilmes@nasa.gov, Robert.E.Wolfe@nasa.gov

The U.S. Global Change Research Program (USGCRP) is a cross-agency coordination of the aggregate research contributions of the U.S. Federal Government to understanding global change. The program is overseen by the White House and has a legislative mandate for periodic reporting about our best scientific assessment.

The USGCRP first deployed the Global Change Information System (GCIS) in conjunction with the release of the Third National Climate Assessment[1], in 2014. While printed and pdf versions of the report were produced, the primary distribution mechanism was a web version hosted at `http://nca2014.globalchange.gov`. The provenance for the report was captured in the GCIS, which itself has a web presence at `http://data.globalchange.gov`. The GCIS assigns globally unique, persistent identifiers to all of the entities, activities and agents relevant to our presentation of that provenance. For the NCA, that includes 43 chapters, 290 figures comprising 495 images, 161 findings, 20 tables and 3,395 references. Many of those (in particular the figures) are linked with activities describing their production and linking to datasets, instruments, models, etc. used to produce them. The structured information describing those items are presented through the web site in a variety of forms, including JSON and Turtle. The semantic web triples are also loaded into a triple store exposed through a SPARQL interface at `http://data.globalchange.gov/sparql`.

More recently, a second major report released by the USGCRP was added to the system in 2016, an assessment of the impacts of climate change on human health.[2]. Again, that report has a web site `http://health2016.globalchange.gov`, and the provenance is distributed through the GCIS at `http://data.globalchange.gov/report/usgcrp-climate-human-health-assessment-2016`. The health report follows the approach taken with the NCA, documenting the 16 chapters, 60 figures comprising 64 images, 29 findings, 4 tables and 1,525 references.

The structured provenance information captured and distributed for those two reports depends heavily on concepts from the W3C PROV Data Model[3], extended with our own GCIS ontology[4].

With two reports now captured within GCIS, more interesting queries now become possible, such as listing agents contributing to both reports, images from each report derived from the same datasets, etc. While the data system capabilities are still in their infancy in terms of impact, we have had the opportunity to describe these first steps of a more formal approach to provenance capture and

representation that we believe lends credence to the findings of the report. These two reports have unprecedented transparency of the entire process and formal documentation of the contributions and documented support for the findings. We are just beginning to link the GCIS artifacts with other data systems and flesh out more complete provenance trails that help tie together our understanding of what we know about global change, and through the provenance how we know those things and where they came from.

References

1. Melillo, J., Richmond, T.T., Yohe, G.: Climate Change Impacts in the United States: The Third National Climate Assessment (2014)
2. Crimmins, A., Balbus, J., Gamble, J.L., Beard, C., Bell, J.E., Dodgen, D., Eisen, R., Fann, N., Hawkins, M., Herring, S.C., Jantarasami, L., Mills, D.M., Saha, S., Sarofim, M., Trtanj, J., Ziska, L.: The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment (2016)
3. Moreau, L., Missier, P.: Prov-dm: The prov data model. Technical report (2012)
4. Ma, X., Tilmes, C., Fox, P.: GCIS ontology version 1.2. (2013)