Ontology engineering for provenance enablement in the third National Climate Assessment

Xiaogang (Marshall) Ma (max7@rpi.edu)[1], Jin Guang Zheng (zhengj3@rpi.edu)[1], Justin Goldstein (jgoldstein@usgcrp.gov)[2,3], Steve Aluenbach (saulenbach@usgcrp.gov)[2,3], Curt Tilmes (Curt.Tilmes@nasa.gov)[3,4], Peter Fox (foxp@rpi.edu)[1]

[1] Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180, USA; [2] University Corporation for Atmospheric Research, Boulder, CO 80301, USA; [3] U.S. Global Change Research Program, Washington, DC 20006, USA; [4] NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

Background and Motivation

Every four years, the U.S. Global Change Research Program (USGCRP) [1] produces a National Climate Assessment (NCA) that presents the findings of global climate change and the impacts of climate change on the United States. The topic of global change builds on a huge collection of scientific research, which also generates provenance information about entities, activities, and people involved in producing datasets, methods and findings. Capturing and presenting global change provenance, linking to the research papers, datasets, models, analyses, observations and satellites, etc. that support the key research findings in this domain can increase understanding, credibility and trust of the assessment process and the resulting report, and aid in reproducibility of results and conclusions.

The USGCRP is now producing the third NCA report (NCA3) and is developing a Global Change Information System (GCIS) that will present the content of that report and its provenance, including the scientific support for the findings of the assessment. As the GCIS will be web-based, it provides a platform for representing the provenance information and implementing the results with semantic web technologies.

Method and Technology

We use a use case-driven iterative development methodology [2] that will present this information both through a human accessible web site as well as a machine readable interface for automated mining of the provenance graph. A use case illustrates an objective that a primary actor wants to accomplish and the sequence of interactions between the primary actor and a system such that the primary actor's objective is successfully achieved. A use case establishes a context in which domain scientists and computer scientists can work together on a topic of interest. Key steps in the iterative methodology are described in Figure 1. Focusing on the technical part, we use the developing World Wide Web Consortium (W3C) PROV data model and ontology [3] for representing the provenance information in the GCIS.

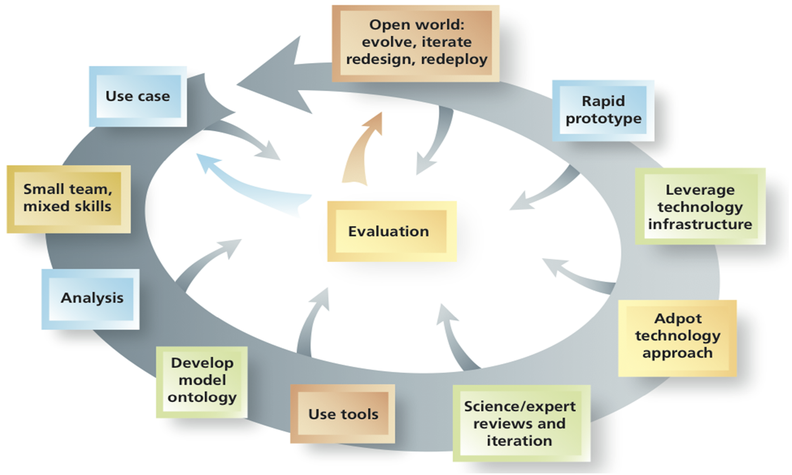

Figure 1 *Semantic Web methodology and technology development process [2]*

Use Case 1: Visit data center website of dataset used to generate a report figure

A viewer wishes to identify the source of the data in a particular NCA3 figure. A reference to the paper in which the figure was originally published in appears in the figure caption. Clicking that reference displays a page of information about the paper, including a link to the datasets used in the paper. Following each of those links presents a page of information about the dataset, including links back to the agency/data center web page describing the dataset in more detail and making the actual data available for order or download.

We collected the primary classes and relationships in this use case (Figure 2) and later adapted them into the GCIS ontology (Figure 5).

Figure 2 *Classes and relationships recognized from Use Case 1*

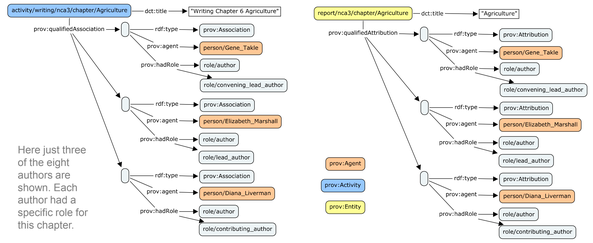

Figure 3 *Roles of people in the writing of chapter 6 (Agriculture) in the NCA3 draft report*

Use Case 2: Roles of people in the generation of a chapter in the NCA3 draft report

A reader sees that Chapter 6 (Agriculture) in the NCA3 draft report was written by a list of authors. On the title page of that chapter the reader can see the role of each author, i.e., convening lead author, lead author or contributing author, in the generation of this report chapter.

To make those roles also machine-readable, we collected classes, instances and relationships in this use case (Figure 3) and adapted them into the GCIS ontology (Figure 5).

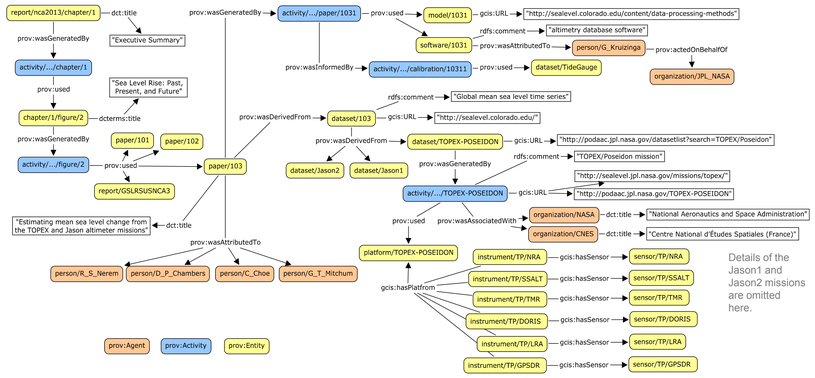

Figure 4 *Platforms, instruments and sensors that contribute to Figure 1.2 in NCA3*

Use Case 3: Provenance tracing of NASA contributions to Figure 1.2 in NCA3 draft report

A reader sees that Figure 1.2 "Sea Level Rise: Past, Present and Future" of the NCA3 draft report cites four data sources in the figure caption. Selecting the third citation displays a page of information about the paper and a citation to the dataset used in the paper. Clicking the citation link the reader opens a page containing information about the dataset, including a description that the dataset is derived from data produced by the TOPEX/Poseidon and Jason altimeter missions funded by NASA and CNES. Following each of these missions presents a page about the platforms, instruments and sensors in that mission.

To make those information both human- and machine-readable, we collected classes, instances and relationships in this use case (Figure 4) and adapted them into the GCIS ontology (Figure 5).

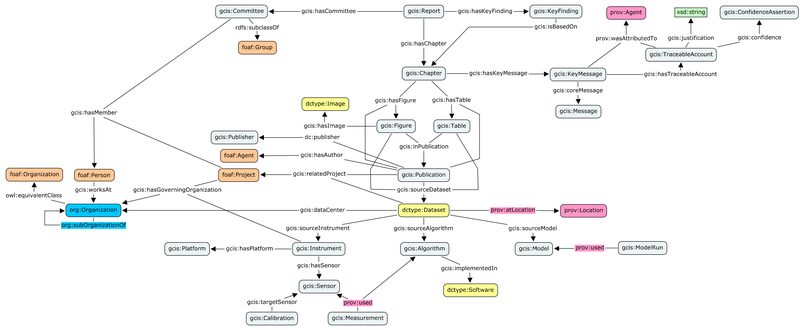

Figure 5 *Primary classes and relationships in current version of the GCIS ontology*

Summary

The ongoing research concentrates on the provenance for the NCA3 report. Following the iterative development methodology, we have worked on a number of use cases to refine an ontology for describing entities, activities, agents and their inter-relationships in the NCA3 report. We also mapped those entities and relationships into the PROV-O ontology to realize the formal presentation of provenance. Several prototype systems have been developed to provide users the functionalities to browse and search provenance information with topics of interest. In the future, the GCIS will collect and link records of publications, datasets, instruments, organizations, methods, people, etc. eventually covering provenance information for the entire scope of global change.

References:

[1] www.globalchange.gov
[2] http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology
[3] http://www.w3.org/TR/2012/WD-prov-overview-20121211

Get the poster at:

Sponsors:

National Science Foundation
University Corporation for Atmospheric Research
Tetherless World Constellation

Acknowledgments:

We thank Stephan Zednik for his contributions to the earlier stage of the GCIS endeavor, and Ana Pinheiro Privette and Anne Waple for their comments on the GCIS.