Tracking and Establishing Provenance of Earth Science Datasets: A NASA-based Example

Hampapuram K. Ramapriyan[1], Justin C. Goldstein[2,3],
Hook Hua[4,5] and Robert E. Wolfe[2,6]

[1] Science Systems and Applications, Inc., Lanham, MD, USA,
Hampapuram.Ramapriyan@ssaihq.com
[2] ICF International, Fairfax, VA USA,
[3] US Global Change Research Program Washington, DC, USA, jgoldstein@usgcrp.gov
[4] California Institute of Technology, Pasadena, CA USA, [5] NASA Jet Propulsion Laboratory, Pasadena, CA USA, hook.hua@jpl.nasa.gov,
[6] NASA Goddard Space Flight Center, Greenbelt, MD USA,
robert.e.wolfe@nasa.gov

Abstract. Information quality is of paramount importance to science. Accurate, scientifically vetted and statistically meaningful and, ideally, reproducible information engenders scientific trust and research opportunities. Therefore, so-called Highly Influential Scientific Assessments (HISA) such as the U.S. Third National Climate Assessment undergo a very rigorous process to ensure transparency and credibility. As an activity to support the transparency of such reports, the U.S. Global Change Research Program has developed the Global Change Information System (GCIS). Specifically related to the transparency of NCA3, a recent activity was carried out to trace the provenance as completely as possible for all figures in the NCA3 report that predominantly used NASA data. This paper discusses lessons learned from this activity that trace the provenance of NASA figures in a major HISA-class pdf report.

Keywords: Information Systems · Data Provenance · Information Quality · HISA Reports · Climate Assessment · Lessons Learned

1 Introduction

Accurate, scientifically vetted and statistically meaningful and, ideally, reproducible scientific information engenders scientific trust and research opportunities. To support the transparency of reports such as the Highly Influential Scientific Assessment (HISA) that is the U. S. Third National Climate Assessment [1], the U.S. Global Change Research Program (USGCRP) has developed the Global Change Information System (GCIS). The GCIS is a web-based resource that facilitates tracing of connections among various entities of which a report is comprised, such as key messages and findings, figures, images used in the figures, data used to generate the images, etc. to foster comprehension of the mechanisms that led to the various conclusions in a report. It is available online at: http://data.globalchange.gov.

This paper presents the lessons learned from a NASA-funded activity to ensure that the NCA3 figure inputs derived from NASA data and their connections to findings and key messages were as complete as possible.

2 GCIS

The GCIS is an open-source, web-based resource for traceable, sound global change data, information, and products. GCIS contains sufficient metadata and links to the sources of data, information and products, guiding users to global change products selected by the 13 USGCRP member agencies. It serves as a key access point to assessments, reports and tools produced by USGCRP [2, 3, 4 5]. The World Wide Web Consortium (W3C) definition of provenance underlies that of GCIS.

The GCIS data model used to structure global change information represents entities such as reports, chapters, figures, bibliographic entries, organizations and people, and uses widely-adopted relationships, including provenance, among such entities. Each item referenced in the GCIS has a unique, persistent identifier takes the form of a Uniform Resource Identifier (URI), but may include other common identifiers such as Universally Unique Identifiers (UUIDs), and Digital Object Identifiers (DOIs).

The W3C Provenance Working Group has defined an interoperable specification (PROV) for the representation of provenance information. The standard is very general, intended to support the breadth of any domain through built in points of extensibility. Generally this provenance can be expressed as {entities (inputs and outputs), agents and activities}. To codify the provenance of GCIS information we are leveraging Provenance for Earth Science (PROV-ES) extension of WC3 PROV that is being developed by the NASA Earth Science Data Systems Working Group (ESDSWG).

We have leveraged the GCIS Application Program Interfaces (API) to extract GCIS content to ingest into a PROV-ES search service for faceted search and provenance exploration. NCA3 content such as figures, persons, and activities are extracted from GCIS as JSON documents. Using scripts, key GCIS concepts and their attributes are mapped onto W3C PROV types. Extensions to baseline PROV concepts are added as additional qualified named attributes. The mapping enables us to map GCIS-specific information into standard and interoperable W3C PROV. For example, a *gcis:Figure* is mapped onto a W3C PROV *prov:Entity*, but with additional attributes.

The GCIS discovery service includes a provenance faceted search capability enabling users to facet navigate GCIS resources in the context of provenance. More specifically, it enables users to “drill-down” by applying a sequential set of selection criteria across different facets (values) of the GCIS content.

3 NCA3 Figures Using NASA Data

NASA data from satellites, instruments and/or models have been used in 20 of the NCA3 figures. One such figure: Figure 1.2 “Flooding and Hurricane Irene”, supports the NCA3 Key Message: “Infrastructure will be increasingly compromised by climate-related hazards, including sea level rise, coastal flooding, and intense precipitation events." It shows an image of Hurricane Irene over the northeastern U.S. acquired from NASA's MODIS instrument on-board the Aqua satellite. The caption was a starting point for gathering more detailed metadata. The figure can be found at http://nca2014.globalchange.gov/report/regions/northeast/graphics/flooding-and-hurricane-irene.

4 Provenance Tracing and Lessons Learned

4.1 Provenance Tracing

The provenance of each of the 20 figures was manually analysed to trace back to the contributing sources (images, data, and analysis methods) to the best extent possible. The results were documented in the form of "activities" used for generating the figures. An activity is defined by a clearly identified set of inputs and outputs, and a method of generating the outputs from the inputs. A majority of the figures require performance of more than one activity. One or more inputs and/or activities may be needed to generate the figure. The method may be as simple as adapting a figure from an article or it could be more complicated and include a detail description of the activity. Where more than one activity is involved in the trace back, activity *n* is used for generating inputs needed for activity *n - 1*. Specification of a complete set of activities for a given figure constitutes its provenance trace.

4.2 Lessons Learned

The key lessons learned from this effort are summarized below. These lessons are similar to the ones reported in [6] regarding experience with collecting metadata for the NCA3 report. They also report that some of the lessons learned have been applied to improve the metadata collection process in the more recent health assessment report planned for release in 2016.

- It is difficult to trace back to derive provenance after reports are completed and delivered. This is because generally, the authors contributing to influential reports are very busy individuals who have spent a considerable amount of time in their research and who have applied significant effort into gathering materials and writing their sections or chapters. If in this process they have not maintained complete documentation to assist in tracing back to derive provenance, then it will involve either more work for the authors or an independent effort to investigate provenance.
- Attempts to follow up with authors on provenance could be misinterpreted as questioning their research.
- To avoid these issues, it is useful to provide the authors with detailed instructions and templates before they start writing their sections or chapters in influential reports. Generally, it is useful to provide readers with information in the form of inputs, outputs and methods (descriptive and/or mathematical) for each dataset used, images or figures generated, and key messages.

- Even with instructions and templates provided to the authors, during the generation of a report it is beneficial for an independent team to check for completeness of traceability from a non-expert reader's point of view. If the independent team is involved starting with the early drafts of the report, the traceability check can be accomplished with minimal impact on the report publication schedule.
- Due to the very nature of a HISA, all underlying information should be held in a long-lived repository and be easily accessible to users for at least as long as the reports are deemed to be of interest to the community.

Acknowledgements

The authors would like to thank the other members of the Provenance Team - Steve Aulenbach, Brian Duggan, Gerald Manipon, Dexter Tan, Curt Tilmes, Brian Wilson and Stephan Zednik. Goldstein's work at the USGCRP National Coordination Office (NCO was supported by an NSF contract with UCAR and a subsequent NASA contract with ICF International. Hua's work was supported by a NASA contract with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA. Ramapriyan's work was supported by a NASA contract with Science Systems and Applications, Inc. Wolfe's work was performed as a NASA employee on detail to USGCRP.

References

1. Melillo, J. M., T.C. Richmond, and G. W. Yohe, Eds. (2014), Climate Change Impacts in the United States: The Third National Climate Assessment. U.S. Global Change Research Program, 841 pp. DOI: 10.7930/J0Z31WJ2.
2. Tilmes, C., P. Fox, X. Ma, D. L. McGuinness, A. P. Privette, A. Smith, A. Waple, S.Zednik, and J. G. Zheng (2013), "Provenance Representation for the National Climate Assessment in the Global Change Information System", IEEE Transactions on Geoscience and Remote Sensing, 51(11), 5160-5168, 2013, DOI: 10.1109/TGRS.2013.2262179.
3. Tilmes, C., A. P. Privette, J. Chen, R. Ramachandran, K. M. Bugbee, and R. E. Wolfe (2015), "Linking from observations to data to actionable science in the climate data initiative", 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 1354-1357, July 2015, DOI: 10.1109/IGARSS.2015.7326027.
4. Waple A. M., S. M. Champion, K. E. Kunkel, C. Tilmes (2016), *Innovations in information management and access for assessments,* Climate Change, Special issue on "The National Climate Assessment: Innovations in Science and Engagement", eds. K. Jacobs, S. Moser, J. Buizer, 1-15, 2016, DOI: 10.1007/s10584-015-1588-7.
5. Wolfe, R.E., Duggan, B., Aulenbach, S.M., Goldstein, J.C., Tilmes, C., and Buddenberg, A. "Providing provenance to instruments through the US global change information system", Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International, 143-145, DOI: 10.1109/IGARSS.2015.7325719.
6. Champion, S. M. and K. E. Kunkel (2015), "Data Management and the National Climate Assessment: A Data Quality Solution", American Geophysical Union, Fall Meeting, December 2015, San Francisco (Presentation charts by personal communication and available on-line at https://goo.gl/aSG4GM).