# Provenance Representation for the National Climate Assessment in the Global Change Information System

Curt Tilmes, *Member, IEEE*, Peter Fox, Xiaogang Ma, Deborah L. McGuinness, Ana Pinheiro Privette, Aaron Smith, Anne Waple, Stephan Zednik, and Jin Guang Zheng

***Abstract*—The important topic of global climate change builds on a huge collection of scientific research. It is common for agencies releasing climate change information to be served with requests for all supporting materials resulting in a particular conclusion. Capturing and presenting global change provenance, linking to the research papers, data sets, models, analyses, observations, satellites, etc., that support the key research findings in this domain can increase understanding and aid in reproducibility of results and conclusions. The U.S. Global Change Research Program is now coordinating the production of a national climate assessment (NCA) that presents our best understanding of global change. We are now developing a global change information system that will present the content of that report and its provenance, including the scientific support for the findings of the assessment. We are using an approach 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. We plan to use the developing World Wide Web Consortium (W3C) PROV data model and ontology for this system. This paper will describe an overview of the process of developing the NCA and how the provenance trail of the report and each of the technical inputs can be captured and represented using the W3C PROV ontology. This will improve the visibility into the assessment process, increase understanding and possibility of reproducibility, and ultimately increase the credibility and trust of the resulting report.**

***Index Terms*—Data management, government information systems, knowledge representation.**

## I. Background

### A. USGCRP

THE U.S. Global Change Research Program (USGCRP) coordinates and integrates federal research on changes in the global environment and their implications for society. The USGCRP began as a presidential initiative in 1989 and was mandated by Congress in the Global Change Research Act (GCRA) of 1990 [1], which called for "*a comprehensive and integrated United States research program which will assist the Nation and the world to understand, assess, predict, and respond to human-induced and natural processes of global change*."

The USGCRP is developing a global change information system (GCIS) that will link to global change information throughout the 13 federal agencies of the program. That information and a depiction of the provenance information and relationships will be presented through the USGCRP Web presence http://globalchange.gov. The first implementation will support the national climate assessment (NCA).

### B. NCA

The GCRA requires a report to the President and the Congress at least every four years that integrates, evaluates, and interprets the findings of the USGCRP; 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; and analyzes current trends in global change, both human-induced and natural, and projects major trends for the subsequent 25 to 100 years.

A federal advisory committee of experts is now producing a report that will become the 2013 NCA and has defined 30 chapters, which are being developed by 62 "convening lead authors" and 180 additional authors. Those designated contributing authors and their institutional affiliations are a critical part of the provenance of the NCA that will be captured. All of that information will not only be part of the printed and Web-based text of the document but also expressed in a formal structured machine-readable format.

Through an open public process, the NCA has received over 500 distinct official technical inputs to the 2013 NCA report, many of which are reports distilling and synthesizing even more information, coming from thousands of individuals around the federal government, nongovernmental organizations, academic institutions, etc. In additional to the official technical inputs, authors are free to use any other sources of relevant information. The inputs used in the 2013 NCA report include peer-reviewed scientific publications, model data, observational

Manuscript received October 15, 2012; revised February 24, 2013; accepted April 8, 2013. Date of publication July 12, 2013; date of current version October 24, 2013.

C. Tilmes is with the NASA Goddard Space Flight Center, Greenbelt, MD 20771 USA (e-mail: Curt.Tilmes@nasa.gov).

P. Fox, X. Ma, D. L. McGuinness, S. Zednik, and J. G. Zheng are with Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180 USA.

A. P. Privette is with North Carolina State University Cooperative Institute for Climate and Satellites, National Oceanic and Atmospheric Administration National Climatic Data Center, Asheville, NC 28801 USA.

A. Smith is with the University Corporation for Atmospheric Research, Boulder, CO 80301 USA.

A. Waple is with the National Oceanic and Atmospheric Administration National Climatic Data Center, Asheville, NC 28801 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2013.2262179

data (physical, societal, and economic), historical data, sectoral and regional assessments, and data at a variety of scales and resolutions. Most original data are archived in long-term agency data centers responsible for long-term stewardship of the items, but some includes "unconventional" information collected from public health departments, states and tribes, nongovernmental organizations, and data collected but not yet reviewed. Where the data are transformed into new graphics, graphs, or charts, the process and methods used must be clearly and reproducibly documented.

This poses a tremendous challenge (and opportunity) for provenance capture, archive, and presentation. We plan to make information about the NCA process itself, as well as the inputs to the process available through a publicly accessible Web site and SPARQL end point. The GCIS will provide links from the content and findings of the NCA back to their antecedent artifacts.

### C. IQA

As a highly influential scientific assessment, the NCA report is subject to the highest standards of the Information Quality Act (IQA) [2]. A scientific assessment is considered "highly influential" if its dissemination could have a potential impact of more than $500 million in any one year on either the public or private sector; if the dissemination is novel, controversial, or precedent-setting; or if it has significant interagency interest. One of the ways that information can exert economic impact is through the costs or benefits of a regulation based on the disseminated information.

The IQA directed the White House Office of Budget and Management (OMB) to issue government-wide guidelines that "provide policy and procedural guidance to federal agencies for ensuring and maximizing the *quality*, *objectivity*, *utility* and *integrity* of information (including statistical information) disseminated by federal agencies" [2]. In those guidelines, OMB defines "quality" as the encompassing term, of which "utility," "objectivity," and "integrity" are the constituents.

In order to provide *objectivity* and *integrity* of content, a strict peer review process has been adopted for the NCA reports. All narrative and images included in a report will be subject to a detailed review process by experts in the scientific community (the National Research Council, among others), by the individual federal agencies involved in the process, by the NCA Development and Advisory Council, and by the public in general. The USGCRP Review and Comment System will be used to accommodate expert and public participation in this process by capturing and organizing comments for the authors and the responses to those comments.

The IQA requirement for *utility* can be better accomplished by creating a transparent system that provides access to information about the processes, analyses, and data sources. A commitment by the NCA to transparency is revealed at different levels: 1) Information sources are captured with full references; 2) quality and provenance of data are clearly described and peer reviewed; 3) data and image details are represented with standardized metadata (ISO 19115); 4) data, documents, and code are version controlled and managed in an accessible configuration management system; 5) narrative assessment and graphics link to source data and information; and 6) the rationale underlying the key messages of the report is captured in *traceable accounts*.

### D. Traceable Accounts

Accompanying most key messages emerging from the NCA 2013 Report, there will be a *traceable account* that summarizes the thought process that leads to a specific narrative. A *traceable account* describes the evidence base and the major uncertainties of a statement and provides an assessment of the confidence, attributed by the authors, based on evidence of agreement in the related communities of expertise. That level of confidence is a qualitative measure (very high, high, medium, and low) specific to each key message. In addition, authors are asked to provide, when possible, a measure of estimated likelihood of impact or consequence following a scale that ranges from "very likely" (greater than nine in ten) to "very unlikely" (less than one in ten). The use of *traceable accounts* will provide access to information concerning the sources of information and the justification supporting the key messages for a given report.

For example, a key message might be something like "The length of the frost-free season (and the corresponding growing season) is increasing nationally, with the largest increases occurring in the western U.S., affecting ecosystems and agriculture. Continued lengthening of the growing seasons across the U.S. is projected."

The traceable account for that message would describe the evidence base supporting that message, citing specific technical inputs. It would also describe new information and remaining uncertainties and include an assessment of the confidence in the message based on the evidence analyzed, for example, "Given the evidence base and remaining uncertainties, confidence is **very high** that the length of the frost-free season (also referred to as the growing season) is increasing nationally, with the largest increases occurring in the western U.S, affecting ecosystems, gardening, and agriculture. Confidence is **very high** that there will be continued lengthening of these seasons across the U.S. given the evidence base."

To increase the general credibility and usability of the NCA report and its content, the NCA aims to go beyond the IQA requirements for transparency and traceability, and support reproducibility for all its content. This will be accomplished by keeping accurate records of all processes and collecting metadata on data sources, images, and code. The objective is to make publicly (and easily) available the information required to allow any user to reproduce that same image independently.

NCA goals for providing reproducibility require a strong commitment from the entire community involved in the process of generating the NCA reports, and the level of effort will require a phased approach to its implementation. Fig. 1 depicts the range of the goals for the NCA, ranging from the basic requirements for traceability through those required for full reproducibility.

By documenting all sources underlying the narrative and images of the report and keeping track of the chain of custody

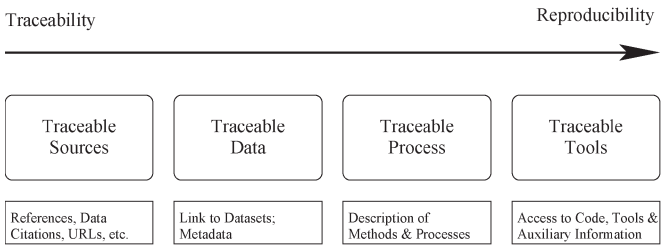


Fig. 1. Transparency goals for the NCA.

for that content, one may ensure compliance with IQA requirements.

## II. Preservation Content

In 1998, the National Aeronautics and Space Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA) sponsored a workshop coordinated through the USGCRP, which produced a report on *global change science requirements for long-term archiving* [3]. It describes a particular content that future consumers of science might be interested in. Building on that and other efforts, the NASA Earth Science Data and Information System project published the NASA Earth Science Data Preservation Content Specification [4] in 2011. It provides guidance on "specific content items that need to be preserved from each of NASA's missions." As the provenance graph from the NCA is explored, figures to papers to datasets, and so forth, the relationships between those items can be followed. Each of the items recommended for long-term preservation and their relationships could be useful for various classes of readers of the NCA. A later example will describe a particular use case in examining the provenance graph describing a representation of the relationships between some of those preservation content items.

### A. Data Citations

One of the major content item relationships of concern is that between a scientific paper and the datasets used by the researcher in producing the conclusions in the paper. New science builds on previous efforts, and citing other papers is a fundamental practice in documenting good science. Documenting data has not always been as rigorously practiced [5]. Recent advances such as the technologies of the World Wide Web have made data easier to obtain and also make it easier to reference and direct others to data that were used. In particular, a study coordinated by Ruth Duerr through the NASA Earth Science Data Systems Working Group and the Federation of Earth Science Information Partners (ESIP) analyzed identification schemes for earth science data and recommended that data centers assign Digital Object Identifiers (DOIs) to datasets just as publishers assign them to journal articles [6]. Guidelines for data citations building on those recommendations were approved by the ESIP Assembly in 2012 [7].

As science journal authors, reviewers and editors start following those guidelines, discovering the relationships between science papers, and their data will become easier not only for direct consumers of their research but also for incorporation of those relationships into citation and provenance graphs in systems such as the GCIS.

For our initial application, the NCA, we have often found data citation practices poor to nonexistent. While some automation can be employed to extract basic information (authors, chapters, key messages, and even reference metadata), a support team is working with the author teams to manually track down input data sources and enter sufficient metadata into the system.

### B. Identifiers

Following the general data model for provenance described in the World Wide Web Consortium (W3C) PROV, described in more detail hereinafter, the GCIS assigns globally unique persistent identifiers to all of the entities, activities, and agents relevant to our discussions of provenance. These allow us to reference the items uniquely, both internally to the GCIS and externally to and from other related systems. We link to existing identifiers where possible and appropriate, using journal- or data-center-assigned DOIs for papers and datasets. The Global Change Master Directory of NASA has also assigned standard reusable identifiers for many of the datasets, services, and other keywords that we are referencing [8]. These include their dataset Entry_ID and universally unique identifiers for keywords.

Each identifier is mapped to a uniform resource identifier (URI) in the USGCRP name space rooted under http://data.globalchange.gov/, allowing use of those identifiers for resource description framework (RDF) and linking with semantic Web and other linked data systems. Our representation of those provenance elements with the PROV ontology is built off of those URIs. A few example URIs are shown in Table I.

All of the URIs will also be resolvable locators through Hypertext Transfer Protocol content negotiation [9] to either human-readable HTML Web pages or machine-readable encodings of the metadata describing the item and linking back to the repository for that item (such as a journal site for a paper or an agency data center for an observational dataset). Where items are derived from or associated with other items, the derived items will link back to the items on which they depend. The "activity" representations associated with the derivation will contain sufficient detail to understand and, ultimately, to reproduce the process that created the derived items.

PROV can be extended with domain-defined types for "agents" like our agency and project organizations and specialized agent roles for the NCA process like the "convening lead authors."

Identifiers for people such as scientific research authors are particularly troublesome, since names can not be relied on as a unique identifier. Several efforts are attempting to address the problem, including ResearcherID[1] and ORCID,[2] and we plan to take advantage of them as they become more widespread. Although we can harvest basic publication metadata from some publisher Web sites, our early efforts with the NCA rely on a somewhat manual (and, we acknowledge, error-prone) process.

[1]http://researcherid.com.
[2]http://orcid.org.

TABLE I
SOME EXAMPLE URIS FOR GCIS

| URI | Resource |
|---|---|
| http://data.globalchange.gov/report/nca2013 | NCA 2013 Report |
| http://data.globalchange.gov/report/nca2013/chapter/13/figure/4 | Figure 4 from Chapter 13 of the report |
| http://data.globalchange.gov/organization/Santa_Clara_University | Santa Clara University |
| http://data.globalchange.gov/person/Nancy_G._Maynard | Convening Lead Author Nancy G. Maynard |
| http://data.globalchange.gov/publication/doi/10.1029/2005jc003136 | Article *Variability of ocean heat uptake: Reconciling observations and models* |

In the system, unique identifiers for authors are recorded where available and assigned where they are not, so we can distinguish multiple authors with the same name within our own database.

### C. Granularity

Earth science data encompass a wide range of data types and forms. Some are as simple as a spreadsheet of *in situ* measurements gathered by hand. Others include millions of measurements from remote sensing earth observation satellite instruments. The organization of such data into meaningful useful portions is often referred to as partitioning or modularizing the data. The NASA Earth Observing System Data Model defines collections of data that are organized into archivable chunks or "granules." For example, the Moderate Resolution Imaging Spectrometer has a granule size of 5 min, with a unique identifier and "granule-level" metadata recorded for each 5-min granule.

When referring to such data, the GCIS will record and link information only at the collection or dataset level, with individual granule level data, metadata, provenance, etc., remaining solely the responsibility of the data center holding that data. In some cases, provenance or lineage information has only been recorded and presented by the data centers at the granule level. It must be summarized to the collection level to be usefully captured and linked into the GCIS. Depending on the processing scenarios, this summarization could be very simple, or it could be quite complicated. Good algorithms for succinctly and correctly determining and representing collection level provenance from granule level processing records are an area that needs further research.

Within the reports such as the NCA, our "granules" will be each identifiable portion of the content where knowledge of particular metadata, including provenance, is important. For the NCA, this will include chapters, which can be linked through attribution to their authors; key messages, which can be linked to their traceable accounts; individual figures; graphs; etc., which can be linked to other research papers or data sets from which they are derived.

Some of the decomposition of the report into those "granules" can be automated by parsing out the structured parts of the document. Some requires manual intervention and "cleanup" to ensure that the document is represented faithfully.

## III. PROVENANCE REPRESENTATION

Many models for provenance representation have been developed in various communities. For the general geoscience provenance and our particular focus for the needs of the GCIS, we have converged on two: the lineage elements for the International Standards Organization (ISO) Geographic information–Metadata standards for specific data production provenance and the W3C PROV standard for representation and interoperability of the overall provenance. These standards overlap in some areas but, in general, are complementary, and each has its uses. In general, we encourage data producers and archives to adopt ISO metadata for complete documentation of their data products and, specifically, to populate the lineage elements with their sources and process for creating those products so that we (and any other users of their products) can discover and use provenance facts about those products. We will use the W3C PROV model for our linking of that data product provenance with all of the other provenance that we are presenting.

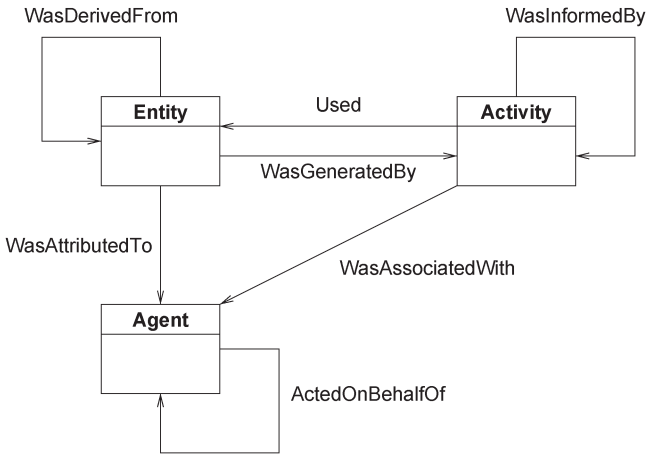


Fig. 2. PROV core structures.

### A. ISO Lineage Elements

Earth science data centers are gradually converging on ISO metadata for representation of basic provenance for their data products. ISO 19115:2003, *Geographic information—Metadata* includes a section `LI_Lineage` with elements `LI_Source` and `LI_ProcessStep` that can represent basic provenance information. ISO 19115-2:2009 extends the data lineage model, allowing more detail (`LE_Source`, `LE_Processing`, `LE_Algorithm`, etc.). Where data producers or archives are describing their data with ISO Extensible Markup Language metadata, we can obtain information about provenance directly from those descriptions and link it into our overall provenance chain.

The description of the provenance of the data products themselves is primarily the responsibility of the data producers and archivers. Unfortunately, few data producers or archives are

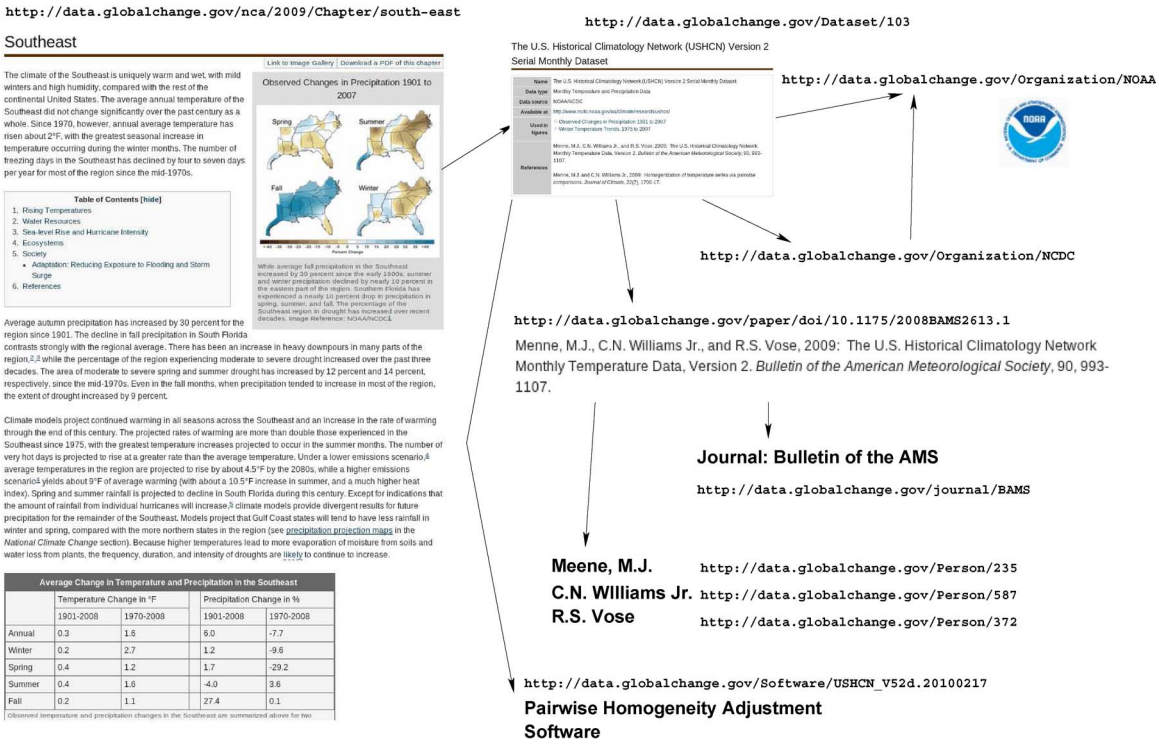


Fig. 3. NCA 2009 figure example provenance.

publishing ISO metadata, and of those even fewer populate the lineage elements with consistent descriptions and identifiers. In this early phase of the GCIS, where our focus is on the primary provenance of the NCA report itself, we are unable to determine and produce complete provenance for data products where the producers have failed to do so. We will be working with producers and archives to encourage good provenance capture and representation using standards like those of the ISO. As they do so, we plan to link their provenance with ours.

### B. W3C PROV

The W3C Provenance Working Group is finalizing definition of the PROV data model (PROV-DM) [10]. PROV-DM is a "generic data model for provenance that allows domain and application specific representations of provenance to be translated into such a data model and *interchanged* between systems." The PROV specification family also includes PROV-O, an OWL2 ontology [11].

The PROV-DM allows some quite complex provenance relationships to be represented. The GCIS, starting with the NCA, will focus on using PROV for some very basic relationships in the short term. These could be extended in the future.

The core types defined by PROV include Entity, Activity, and Agent. Those types and the most basic relationships between them are depicted in Fig. 2. Early representations of the NCA provenance will use the most important and useful PROV concepts: Attribution ("WasAttributedTo") and Derivation ("WasDerivedFrom"). Those simple concepts alone are sufficient to build a graph of contributing entities and agents and construct a Web site for a user to browse the collection.

Where appropriate and useful, we will extend those by further describing the activities involved and attaching additional attributes to the relationships between entities, activities, and agents.

## IV. NCA 2009 W3C PROV-O Example

In order to express the functionality of the GCIS in terms of how the user will interact with it for provenance information, we applied a number of use cases. In this use case, we will describe an application of provenance in the GCIS, with a focus on the NCA report, shown with a simple example from the NCA 2009 report depicted in Fig. 3.

The reader of the NCA report wishes to identify the dataset used to generate a particular figure in the report. S/he is directed first to the figure caption. Selecting the caption displays a page of information about the figure, and if the figure was originally published in another paper, it includes a link via the paper's DOI to the publisher's site describing that paper, offering it for download. The page of information also includes references to the datasets used in the paper on which the figure was based. Following each of the dataset links presents a page of information about the dataset, including links back to the agency/data center Web page which provides more details on the dataset (metadata) and from which the actual data may be available for order or download.

More formally, the primary *actor* in this use case is a reader of the NCA report. The *precondition* is that the reader is browsing the NCA report online, and after consulting the basic flow of information described previously, the *postcondition* is that the reader has visited the dataset Web page on the data center Web site. The components in the use case include a report, a chapter, a figure, a paper, and a dataset.

We captured provenance information required to support the aforementioned use case and modeled it using the PROV-O

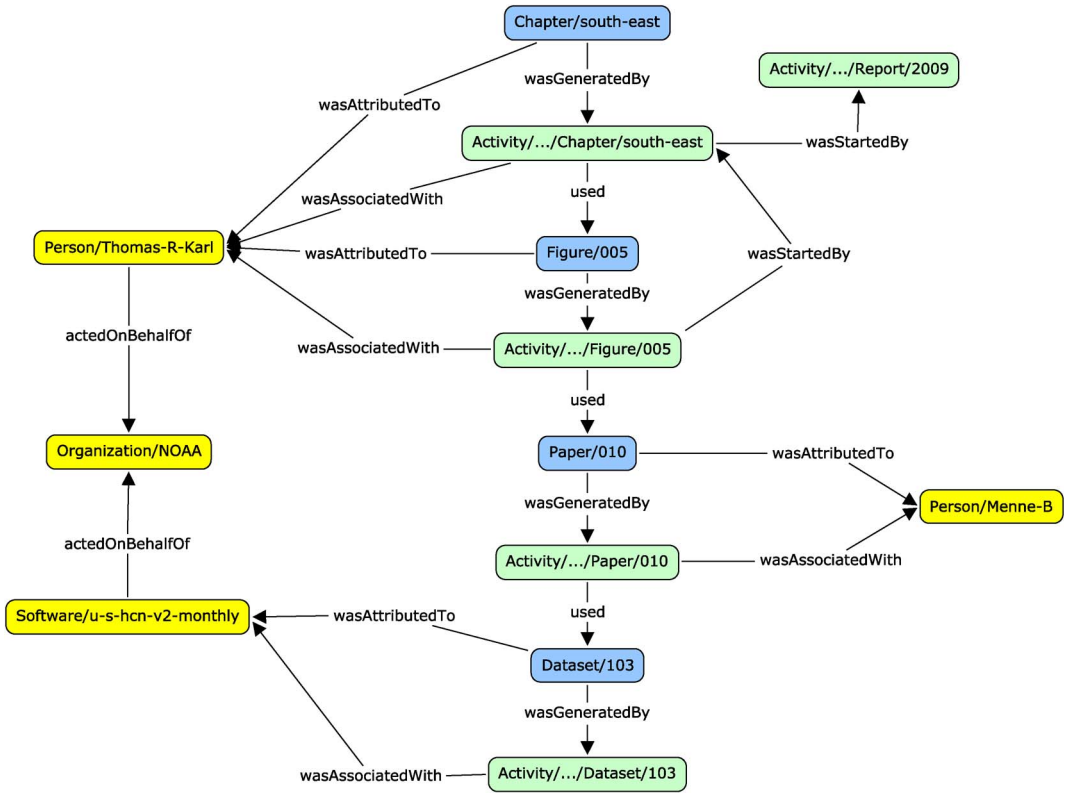


Fig. 4. PROV-O example graph representation.

ontology [11]. The example in Fig. 4 illustrates the provenance history of a figure in a chapter of the NCA report. This fulfills the requirements of the aforementioned use case. Aside from chapter, figure, paper, and dataset, we also captured provenance information such as persons involved in generating a chapter, paper, or figure, software used to collect the dataset, etc. Fig. 4 applies the three starting point classes (i.e., Entity, Activity, and Agent) in the PROV-O ontology to depict the provenance information. For example, the entity "Chapter south-east" (a chapter) was generated by the activity "Chapter south-east," which used another entity "Figure 005." Entities "Chapter south-east" and "Figure 005" were attributed to the agent "Person Thomas-R-Karl," and the activity "Chapter south-east" was associated to the same agent. Another PROV relationship "wasStartedBy" was used to indicate that the activity of producing "Figure 005" was started by the activity of creating the "Chapter south-east," which, in turn, was started by the activity of creating the whole report. Fig. 5 shows an RDF serialization of the resources depicted in Fig. 4. The components shown in the provenance history also reflect the aforementioned consideration on the granularity of data and information in GCIS. This example shows that data provenance and data transparency can be easily captured and supported by adopting the PROV-O ontology.

## V. Conclusion and Future Work

Using the semantic Web approach to organizing and presenting the provenance of the NCA through the GCIS has resulted in a more structured and interoperable representation, allowing other tools and services to interact with the data behind the Web site, increasing its usefulness for other purposes, including data mining and service integration beyond what is possible with a simple Web site. This is a significant feature of the semantic Web and/or linked data approach.

The current focus for the GCIS in support of the NCA is "simple" provenance. A reader of the report will be able to read the content of the report, including a prose description of the rationale for the key messages of the report. From that, the reader can explore the papers, datasets, etc., that contributed to those descriptions and conclusions.

This approach initially uses relatively lightweight annotations and associated reasoning. Some of the authors have created and used a conversion tool to take in data and create lightly annotated data automatically [12]. This can be done with little effort since it is an automated effort. Additional tools are in process that support enhancements with minimal human effort.

Starting with the core of entities represented in support of the NCA, the GCIS will build its database of papers, datasets, etc., eventually covering the entire scope of global change. By linking those items through provenance and other relationships, the resulting database could support more complicated data mining beyond the simple walking step by step through the graph on a Web site. It could, for example, produce metrics such as the number of papers that use data derived from each of the Earth observing satellites. It could provide the basis for discovery of interagency dependences such as an Environmental Protection Agency analysis that used data from a Department of Energy model that used data from a NOAA geophysical dataset derived from observations from a NASA satellite. It can be very difficult

```
<http://data.globalchange.gov/nca/2009/Chapter/south-east>
 a prov:Entity;
 prov:wasGeneratedBy <http://data.globalchange.gov/Activity/Writing/nca/2009/Chapter/south-east>;
 prov:wasAttributedTo <http://data.globalchange.gov/Person/Thomas-R-Karl>;
.

<http://data.globalchange.gov/Activity/Writing/nca/2009/Chapter/south-east>
 a prov:Activity;
 prov:used <http://data.globalchange.gov/Figure/005>;
 prov:wasAssociatedWith <http://data.globalchange.gov/Person/Thomas-R-Karl>;
 prov:wasStartedByActivity <http://data.globalchange.gov/Activity/Writing/nca/2009/report>;
.

<http://data.globalchange.gov/Figure/005>
 a prov:Entity;
 prov:wasGeneratedBy <http://data.globalchange.gov/Activity/Writing/Figure/005>;
 prov:wasAttributedTo <http://data.globalchange.gov/Person/Thomas-R-Karl>;
.
<http://data.globalchange.gov/Activity/Writing/Figure/005>
 a prov:Activity;
 prov:used <http://data.globalchange.gov/Paper/010>;
 prov:wasAssociatedWith <http://data.globalchange.gov/Person/Thomas-R-Karl>;
 prov:wasStartedByActivity <http://data.globalchange.gov/Activity/Writing/nca/2009/Chapter/south-east>;
.

<http://data.globalchange.gov/Paper/010>
 a prov:Entity;
 prov:wasAttributedTo <http://data.globalchange.gov/Person/Menne-B>;
 prov:wasGeneratedBy <http://data.globalchange.gov/Activity/Writing/Paper/010>;
.

<http://data.globalchange.gov/Activity/Writing/Paper/010>
 a prov:Activity;
 prov:used <http://data.globalchange.gov/Dataset/103>;
 prov:wasAssociatedWith <http://data.globalchange.gov/Person/Menne-B>;
.

<http://data.globalchange.gov/Dataset/103>
 a prov:Entity;
 prov:wasGeneratedBy <http://data.globalchange.gov/Activity/Generating/Dataset/103>;
 rdfs:seeAlso <http://www.ncdc.noaa.gov/oa/climate/research/ushcn/>;
 prov:wasAttributedTo <http://data.globalchange.gov/Software/u-s-hcn-v2-monthly>;
.

<http://data.globalchange.gov/Activity/Generating/Dataset/103>
 a prov:Activity;
 prov:wasAssociatedWith <http://data.globalchange.gov/Software/u-s-hcn-v2-monthly>;
.
```

Fig. 5. PROV-O example turtle representation.

and manpower intensive to manually trace such relationships in the absence of a standard representation of provenance crossing those interagency boundaries.

In future work in the GCIS and for the continuing sustained assessment activities, we plan to capture more details concerning rationales for decisions. We are moving toward more descriptive traceable accounts that capture not only the basic sources of data and analyzers but also more details of how the analysis progressed along with its assumptions and manipulations. We are exploring how this can be captured using the the Proof Markup Language (PML) [13]. PML was initially designed to provide a formal foundation for capturing information concerning basic provenance, information manipulations, and trust in a unified formal framework. It was one of the provenance languages that influenced today's W3C proposal for a provenance language recommendation. The next generation of PML provides compatibility with PROV and extends it to support encodings of more detailed information manipulation steps such as those used in coming to recommendations in a global change report. The next generation of traceable accounts would include formal encoding of the rationale and would not just be consumable by humans but they could also be analyzed mechanically for correctness and consistency.

## ACKNOWLEDGMENT

The authors would like to thank the staff and management of the U.S. Global Change Research Program National Coordination Office and the National Climate Assessment. This work builds on years of discussions within those groups, NASA's Earth Science Data Systems Working Group, and the Federation of Earth Science Information Partners.

## REFERENCES

[1] *U.S. Code Global Change Research Act of 1990 (P.L. 101–606)*, U.S. House of Representatives, Washington, DC, USA, 1990.

[2] Section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001 (Public Law 106-554). Federal Register, vol. 67, no. 36.

[3] "Global change science requirements for long-term archiving, report of the workshop," Washington, DC, USA, Tech. Rep., Oct. 28–30, 1998, Sponsored by NASA and NOAA, through the USGCRP Program Office.

[4] H. H. K. Ramapriyan and J. F. Moses, "NASA Earth science data preservation content specification," Goddard Space Flight Center, Greenbelt, MD, USA, Tech. Rep., 2011, Earth Science Data and Information System Project, Code 423.

[5] M. A. Parsons, R. Duerr, and J.-B. Minster, "Data citation and peer review," *EOS, Trans. AGU*, vol. 91, no. 34, pp. 297–298, Aug. 2010.

[6] R. Duerr, R. Downs, C. Tilmes, B. Barkstrom, W. Lenhardt, J. Glassy, L. Bermudez, and P. Slaughter, "On the utility of identification schemes for digital Earth science data: An assessment and recommendations," *Earth Sci. Inf.*, vol. 4, no. 3, pp. 139–160, Sep. 2011.

[7] D. S. Committee. (2012). Data citation guidelines for data providers and archives, Federation of Earth Science Information Partners, Raleigh, NC, USA, Tech. Rep. [Online]. Available: http://commons.esipfed.org/node/308
[8] L. M. Olsen, G. Major, K. Shein, J. Scialdone, S. Ritz, T. Stevens, M. Morahan, A. Aleman, R. Vogel, S. Leicester, H. Weir, M. Meaux, S. Grebas, C. Solomon, M. Holland, T. Northcutt, R. A. Restrepo, and R. Bilodeau, 2013, NASA/Global Change Master Directory (GCMD) Earth Science Keywords. Version 8.0.0.0.0.
[9] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, Hypertext Transfer Protocol—HTTP/1.1, The Internet Engineering Task Force (IETF), Fremont, CA, USA, IETF, RFC 2616. [Online]. Available: http://www.ietf.org/rfc/rfc2616.txt
[10] K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, and C. Tilmes, Prov-DM: The Prov data model, W3C/MIT/CSAIL, Cambridge, MA, USA, Tech. Rep. [Online]. Available: http://www.w3.org/TR/prov-dm/
[11] K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao, Prov-o: The Prov ontology, W3C/MIT/CSAIL, Cambridge, MA, USA, Tech. Rep. [Online]. Available: http://www.w3.org/TR/prov-o/
[12] T. Lebo, J. Erickson, L. Ding, A. Graves, G. Williams, D. DiFranzo, X. Li, J. Michaelis, J. Zheng, J. Flores, Z. Shangguan, D. McGuinness, and J. Hendler, "Producing and using linked open government data in the TWC LOGD portal,"
[13] P. P. da Silva, D. L. McGuinness, and R. Fikes, "A Proof Markup Language for semantic Web services," *Inf. Syst.*, vol. 31, no. 4/5, pp. 381–395, Jun. 2006.

**Curt Tilmes** (M'06) received the B.S. degrees in computer engineering and computer science from Virginia Tech, Blacksburg, VA, USA, in 1991, the M.S. degree in computer science and the M.S. degree in system engineering with a project management concentration from Johns Hopkins University, Baltimore, MD, USA, in 1997 and 2002, respectively, and the Ph.D. degree in computer science from the University of Maryland, Baltimore County, MD, USA, in 2011.

He is currently a Computer Scientist with the Goddard Space Flight Center, National Aeronautics and Space Administration, Greenbelt, MD, USA. He has contributed to several data processing systems for Earth climate monitoring satellite projects, including the Moderate Resolution Imaging Spectrometer and the Ozone Monitoring Instrument. He is on a detail to the U.S. Global Change Research Program as the Technical Lead for the Global Change Information System.

**Peter Fox** received the B.Sc.(Hons.) degree and the Ph.D. degree in applied mathematics (physics and computer science) from Monash University, Melbourne, Australia.

He is currently with Rensselaer Polytechnic Institute, Troy, NY, USA, where he is the Tetherless World Constellation Chair, a Professor of Earth and environmental science and computer science, and the Director of the Information Technology and Web Science Program. His research covers the fields of solar and solar-terrestrial physics, ocean and environmental informatics, computational and computer science, and distributed semantic data frameworks. The results are applied to large-scale distributed data science investigations.

Prof. Fox is the Chair of the International Union of Geodesy and Geophysics Union Commission on Data and Information and serves on the editorial boards of many prominent Earth and space science informatics journals. He was the recipient of the 2012 European Geoscience Union Ian McHarg/Earth and Space Science Informatics Medal and the Earth Science Information Partner's Martha Maiden Lifetime Achievement award for service to the Earth science information communities.

**Xiaogang Ma** received the B.Eng. degree in land resources management and the D.Eng. degree in geoinformatics engineering from the China University of Geosciences, Wuhan, China, in 2002 and 2009, respectively, and the Ph.D. degree from Twente University, Enschede, The Netherlands, in 2011.

Since 2012, he has been a Postdoctoral Research Associate with Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA, with a focus of research on semantic eGeoscience. Recently, he has contributed to the information modeling and semantic application prototype project of the Global Change Information System and the data science project for the Deep Carbon Observatory. His research interests include geothesauri, geo-ontologies, geodata interoperability, geoconceptual modeling, data visualization, and geodata services with the World Wide Web Consortium and Open Geospatial Consortium standards.

**Deborah L. McGuinness** received the B.S. degree in math and computer science from Duke University, Durham, NC, USA, the masters degree in computer science from the University of California, Berkeley, CA, USA, and the Ph.D. degree in computer science from Rutgers University, New Brunswick, NJ, USA.

She is currently with Rensselaer Polytechnic Institute, Troy, NY, USA, where she is the Tetherless World Senior Constellation Chair, a Professor of computer and cognitive science, and the Founding Director of the Web Science Research Center. She is a leading authority on the semantic Web and has been working on knowledge representation and reasoning environments for over 25 years. She also founded McGuinness Associates that consults on semantic applications in a wide range or areas with recent focus on health and environmental informatics, context-aware mobile computing, and next-generation journalism. She leads active research efforts in explanation, trust, ontology environments, and provenance. She is also known for semantic application environments, particularly for eScience frameworks such as the semantic eScience framework and demonstration portals including many in natural science and health informatics settings. Her primary research focuses on making smart systems understandable and usable by a broad range of people.

**Ana Pinheiro Privette** received the B.S. degree from the New University of Lisbon, Lisbon, Portugal, in 1990, the M.Eng. degree in civil and environmental engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1998, and the Ph.D. degree in environmental engineering from the New University of Lisbon in 2003.

She conducted research on remote sensing of land surface temperature (LST), investigating the optimal retrieval of LST through product development (the Advanced Very High Resolution Radiometer and the Moderate Resolution Imaging Spectrometer), modeling, and *in situ* validation efforts. She spent most of her career at the Goddard Space Flight Center, National Aeronautics and Space Administration, and the National Oceanic and Atmospheric Administration National Climatic Data Center (NOAA NCDC), Asheville, NC, USA. She has recently joined the National Climate Assessment Technical Support Unit, North Carolina State University Cooperative Institute for Climate and Satellites, NOAA NCDC, as the National Climate Assessment (NCA) Data Coordinator leading the efforts to create a data management strategy, and a data policy, for the NCA and the U.S. Global Change Research Program Global Change Information System.

**Aaron Smith** received the B.A. degree in international studies and philosophy from American University, Washington DC, USA, and the Master of Public Administration and M.A. degree in international relations, Syracuse University, Syracuse, NY, USA.

He is currently with the University Corporation for Atmospheric Research, Boulder, CO, USA. He is responsible for the Web system development and administration for the constellation of the LAMP-based U.S. Global Change Research Program (USGCRP) Web infrastructure, including developing, designing, maintaining, and optimizing secure and scalable Web applications for the program's Web site. He is the point of contact for the USGCRP National Coordination Office's network, with oversight of contracted system administration, and performs computer maintenance, including server- and client-side Web applications, requiring competency in Windows, Mac, and Linux operating systems. He also gathers configuration requirements, designs, codes, and tests independently, as well as leads the Web development team.

**Anne Waple** received the B.Sc. degree from the University of Wales, Swansea, U.K., and the Ph.D. degree from the University of Massachusetts, Amherst, MA, USA.

She is currently with the National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center, Asheville, NC, USA. She holds joint appointments with the U.S. Global Change Research Program (USGCRP) and with Second Nature. For the USGCRP, she works on incorporating the national climate assessment (NCA) into the Global Change Information System and on presenting the material in a usable and accessible manner. For Second Nature, she is helping to develop a program that will assist higher education institutions incorporate climate adaptation and resilience into their planning, curricula, research, and community programs. She worked in association with NOAA for around 11 years and was most recently the Program Manager for the Assessment Services Program of NOAA and the Chair of the NCA Technical Support Unit.

**Stephan Zednik** received the B.S. degree in computer science from the University of Colorado, Boulder, CO, USA.

Since December 2008, he has been a Software Engineer with Tetherless World Constellation, Rensselaer Polytechnic Institute (RPI), Troy, NY, USA. Prior to coming to RPI, he was with the High Altitude Observatory, National Center for Atmospheric Research, Boulder, CO, USA. He served on the W3C Provenance Working Group, contributing to the PROV recommendations, and was an Editor for PROV-Extensible Markup Language. He led the initial information modeling and ontology development for the Global Change Information System. He has since become a Developer for IQNavigator.

**Jin Guang Zheng** received the B.S. degree in computer science from Rensselaer Polytechnic Institute, Troy, NY, USA, in 2009, where he is currently a fourth-year Ph.D student in computer science and is working with Prof. Fox and Dr. Ma.

Recently, he has contributed to the information modeling and application prototype development in Global Change Information System project. His research interests include using semantic Web technologies to solve problems in scientific domains. He is also interested in semantic similarity computation on the semantic Web and related applications.