Providing global change information for decision-making: capturing and presenting provenance

Xiaogang Ma[1]*, Peter Fox[1], Curt Tilmes[2,3], Katharine Jacobs[4,5] and Anne Waple[6,7]

[1]Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA

[2]NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA

[3]U.S. Global Change Research Program, Suite 250, 1717 Pennsylvania Avenue, NW, Washington, DC 20006, USA

[4]Center for Climate Adaptation Science and Solutions, Institute of the Environment, University of Arizona, Tucson, AZ 85721, USA

[5]Department of Soil, Water and Environmental Science, University of Arizona, Tucson, AZ 85721, USA

[6]Second Nature Inc., Boston, MA 02108, USA

[7]Waple Research and Consulting, Weaverville, NC 28787, USA

*e-mail: max7@rpi.edu

Short Summary: Global change information demands access to data sources and well-documented provenance to provide evidence needed to build confidence in scientific conclusions and, in specific applications, to ensure the information's suitability for use in decision-making. A new generation of Web technology, the Semantic Web, provides tools for that purpose.

The topic of global change covers changes in the global environment (including alterations in climate, land productivity, oceans or other water resources, atmospheric composition and/or chemistry, and ecological systems) that may alter the capacity of the Earth to sustain life and support human systems [1]. Data and findings associated with global change research are of great public, government, and academic

concern and are used in policy and decision-making, which makes the provenance of global change information especially important. In addition, since different types of decisions benefit from different types of information, understanding how to capture and present the provenance of global change information is becoming more of an imperative in adaptive planning.

Provenance tracking

Many science issues in global change research are complex and yield multiple implications for management, such as sea level rise projections. In some cases, including in academic research and for broad scale policy discussion, it is useful to examine probabilistic estimates of future sea level rise (e.g. what is the 'most likely' amount of change over time), whereas in other instances, it is helpful to refer to a broader range of plausible sea level changes (e.g. risk-based framing for a wide variety of coastal management decisions). In 2013 the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC) [2] examined four different greenhouse gas emissions pathways and associated projections of sea level rise. They concluded that it was 'likely' that global sea levels would be between 0.26 to 0.81 meters above recent levels by the last two decades of current century. As a comparison, a report published in 2012 by the National Oceanic and Atmospheric Administration (NOAA)'s Climate Program Office (CPO) [3] in support of the National Climate Assessment [4], provided a wider range of plausible sea level rise describing scenarios from 0.2 to 2.0 meters above 1992 levels by 2100.

The above two reports took different approaches for different reasons; the CPO report included higher and lower estimates of sea level rise because they were considered useful in risk-based decision-making even though the IPCC considered them less 'likely'. In other words, awareness of low probability and high impact futures is useful for coastal decision-making, but arguably less valuable for broad scale policy framing. For limiting risk to life and property, it is therefore useful to understand the nature of the scenarios and projections, especially at the higher and lower end of the range. In order to do this it is

necessary to know what data, model, and analytical sources are used in each report and to be able to access them easily and in an interpretable format — a goal that provenance tracking can support.

The literal meaning of provenance is the origin or source of something. As discussed above, understanding the origins of global change information is not only critical to people's understanding of how to use it, but also facilitates integration of knowledge across disciplines. This kind of interdisciplinary assessment is important in managing the risks associated with global change. For the above example, the AR5 and the CPO report provided textual descriptions for tracking the provenance of their projections but the descriptions are relatively technical and not easy to interpret in terms of what is 'best' or most appropriate in any given circumstance. A more useful approach, in the long term, is to provide search options that facilitate access to data and allow traceability back to original sources, authors, programs and even observing systems.

Although provenance tracking is useful and important, as shown by the above example, capturing and presenting provenance is not an easy task. In scientific works documenting provenance includes linking a range of observations and model output, research activities, people and organizations involved in the production of scientific findings with the supporting datasets and methods used to generate them [5]. It requires significant effort to identify, extract, link, and assemble pieces of information from accumulated documents, codes, datasets, and so on.

Categorization and annotation

The root of provenance capture grows from the soil of metadata collection. Metadata are data about data. Traditionally seen in library catalogs, metadata have received significant attention in the last decade and several metadata standards have been developed to address the recent data deluge challenge. One widely used standard is the Dublin Core metadata schema [6], which consists of 15 core elements, such as title, creator, subject, description, publisher, and so on, that can be used to describe a resource. The schema is

simple, which is an advantage for use. However, it is also a weakness because the schema does not define categories of resources and the provided core elements do not accommodate rich annotation of a resource, such as the geographical and temporal location or resolution of a dataset that could be useful for narrowing down its applicability. Even with the extended metadata terms [7] released by the Dublin Core Metadata Initiative (DCMI) — an open organization supporting shared innovation in metadata design and best practices — such an issue still exists. For example, the resource type 'Bibliographic Resource' in Fig. 1 is a term in the DCMI extended terms [7], whereas the specific type 'Figure' is not and has to be added by the metadata curator.

[Insert Figure 1 about here]

Other metadata standards further extend the functionality for annotating the production process of a resource, such as the data lineage model in ISO 19115 [8] and ISO 19115-2 [9]. Two initiatives, the Proof Markup Language (PML) [10] and the Open Provenance Model (OPM) [11], have developed enriched categories and properties for representing and capturing provenance. Three top categories — entity, process and agent — arise from those works. By using them one can describe a process such as the generation of a figure through which an entity (the figure) was produced, the source entities such as datasets and models used in the process, and agents such as people and/or organizations participated in the process. Those categories became the core of the recent World Wide Web Consortium (W3C) PROV Data Model (PROV-DM) [12] that was derived from those earlier efforts.

Another significant advancement in those metadata schemas and provenance models [7, 10–12] is building on the principles of linked data, using Uniform Resource Identifiers (URIs) as identifiers of resources, rather than literal values. For example, the third reference record in Fig. 1 may be replaced by a URI 'http://data.globalchange.gov/article/10.1080/01490419.2010.491031' that points to a webpage presenting more information about that document, such as title, document type, source journal,

publication year, and so on. The use of URI is just one of the many features that are enabled by the Semantic Web, a new generation of the World Wide Web [13]. The Semantic Web presents a Web of Data compared to the traditional Web of Documents. It adds machine-readable meanings such as more specifically defined categories and annotations to data by using ontologies — specifications of concepts and relations among them — and vocabularies — arranged terms of certain topics — encoded in the Resource Description Framework (RDF) format [14]. For instance, the webpage corresponding to the above URI is just the front end of a piece of data stored in an online database. While humans read the textual description on the webpage and know that it is about a journal article, machines can read data from the back end and recognize the resource type 'bibo:Article' (for meaning of the prefix 'bibo' see Table 1) by tracking the ontologies used in the data.

[Insert Table 1 about here]

Linking for tracking

The aforementioned categorization and annotation focus on the description of individual entities, processes and agents. The Semantic Web allows links to be established among those individual instances, such that in provenance tracking one can retrieve not only the literal description of a dataset used in a research but also an accessible or downloadable version of the dataset itself. In the Global Change Information System (GCIS) [15] under development through the U.S. Global Change Research Program (USGCRP), those Semantic Web technologies have been utilized to capture and present provenance information. The initial focus of GCIS is to support the third United States National Climate Assessment (NCA3). It will present the NCA3 report and also incorporate integrated access to inter-linked resources supporting that report. This significantly enhances transparency of the report, and also enhances the ability of decision-makers to understand the conclusions and to use the underlying data for their own purposes.

The inter-links among provenance information in the GCIS are realized by using properties defined in a number of ontologies, including the PROV Ontology (PROV-O) [16] — an ontology for representing and interchanging provenance information from the W3C PROV-DM. For example, Fig. 2 shows a part of the provenance information captured from the third reference in Fig. 1. Most properties in that graph are from the PROV-O. Properties from a few other domain-specific ontologies, such as the Bibliographic Ontology and the GCIS Ontology (Table 1), are also used. Those specific properties better describe a few relationships — such as those between instruments and sensors — than the general properties from the PROV-O.

[Insert Figure 2 about here]

By categorizing, annotating and linking provenance information, the finished GCIS will be capable of answering provenance-tracking questions for the final NCA3, such as (1) which datasets were used in the analysis and projection of global sea level rise or (2) which agencies and individuals are working on projects related to societal impacts of extreme weather events. The GCIS is intended to be a Web-based source of authoritative, accessible, usable and timely information about global change for use by scientists, decision-makers, and the public. The Semantic Web technology will help make the GCIS a part of the Web of Data, such that other tools and services are also able to interact with data and information in the GCIS and thus create added values in approaches that are applied to address socio-economic, physical, ecological and other intellectual challenges.

Persistent and universally resolvable identifiers such as DOI (Digital Object Identifier) are widely accepted for research articles and increasingly also for data; ResearcherID and ORCID (Open Researcher & Contributor ID) for people makes literature and data easily accessible and citable; and global change research increasingly benefits from the open-access literature and datasets [17]. We argue that the global change research community should take one step further with the curation of provenance information —

such as what is going on with the GCIS. Those works as a whole promote the meaningful eScience [18] — digital or electronic facilitation of science — and wider participation from the global change research community is desired.

Concluding thoughts

As global change information becomes both more abundant and increasingly critical, our need to know more about what, how, when, where, and why information is produced is becoming ever more necessary. Well-curated provenance information makes scientific workflows transparent and improves the credibility and trustworthiness of their outputs. It also facilitates informed and rational policy and decision-making based on the outputs of global change research. For all these reasons, the work on provenance is timely and foundational and is now an embedded component of the Global Change Information System, and a sustained approach to climate assessment.

References

[1] U.S. Code *Global Change Research Act of 1990 (Public Law 101-606)* (U.S. Code, 1990).

[2] Stocker, T. F. *et al.* (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the fifth Assessment Report of the Intergovernmental Panel on Climate Change. Summary for Policymakers* (IPCC, 2013). Available via: http://www.climatechange2013.org/images/uploads/WGI_AR5_SPM_brochure.pdf

[3] Parris, A. *et al. Global Sea Level Rise Scenarios for the United States National Climate Assessment. NOAA Tech Memo OAR CPO-1* (NOAA's Climate Program Office, 2012).

[4] http://assessment.globalchange.gov

[5] Tilmes, C. *et al. IEEE T. Geosci. Remote* **51,** 5160–5168 (2013).

[6] ISO *ISO 15836: Information and Documentation — The Dublin Core Metadata Element Set* (ISO, 2003).

[7] DCMI Usage Board *DCMI Metadata Terms* (DCMI Usage Board, 2012). Available via: http://dublincore.org/documents/dcmi-terms

[8] ISO *ISO 19115: Geographic Information — Metadata* (ISO, 2003).

[9] ISO *ISO 19115-2: Geographic Information — Metadata — Part 2: Extensions for Imagery and Gridded Data* (ISO, 2009).

[10] Moreau, L. *et al. Future Gener. Comp. Sys.* **27,** 743–756 (2011).

[11] Pinheiro da Silva, P., McGuinness, D. L. & Fikes, R. *Inform. Syst.* **31,** 381–395 (2006).

[12] Moreau, L. & Missier, P. (eds.) *PROV-DM: The PROV Data Model* (2013). Available via: http://www.w3.org/TR/prov-dm

[13] Hendler, J. *Science* **299,** 520–521 (2003).

[14] RDF Working Group *Resource Description Framework (RDF)* (W3C, 2004). Available via: http://www.w3.org/RDF/

[15] USGCRP *Global Change Information System Facts Sheet* (USGCRP, 2012). Available via: http://downloads.globalchange.gov/factsheets/gcis_factsheet.pdf

[16] Lebo, T., Sahoo, S. & McGuinness, D. L. (eds.) *PROV-O: The PROV Ontology* (2013). Available via: http://www.w3.org/TR/prov-o

[17] Overpeck, J. T., Meehl, G. A., Bony, S. & Easterling, D. R. *Science* **331,** 700–702 (2011).

[18] Fox, P. & Hendler, J. in *The Fourth Paradigm: Data-Intensive Scientific Discovery* (eds Hey, T., Tansley, S. & Tolle, K.) 147–152 (Microsoft Res., 2009)

Acknowledgements

This work was supported by the National Science Foundation grant through the University Corporation for Atmospheric Research to Rensselaer Polytechnic Institute under contract S13-94358. The authors thank Jin Guang Zheng, Justin Goldstein, Linyun Fu, Patrick West, Steven Aulenbach, Stephan Zednik, and Brian Duggan for collaborations in the research project Global Change Information System:

Information Model and Semantic Application Prototypes (GCIS-IMSAP), and Tim Lebo and Deborah McGuinness for their comments on PROV-O, OPM and PML.

Author contributions

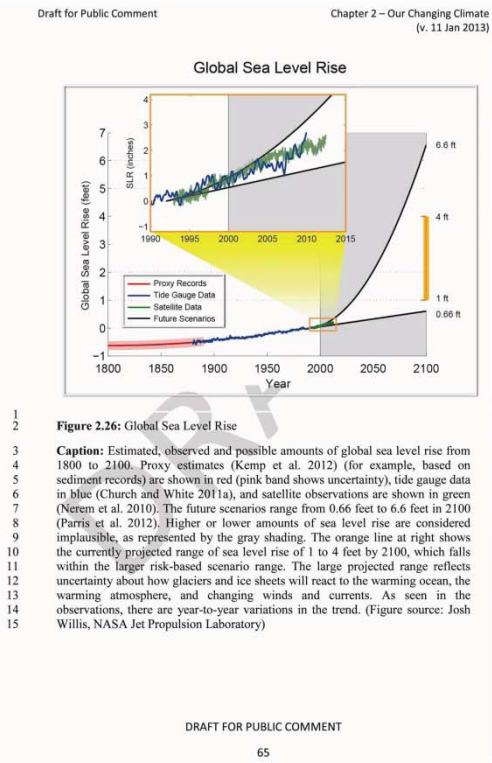
All authors contributed to the planning of the paper. X.M. led the work. P.F. contributed the use case-driven Semantic Web application method which generated the provenance graph in the global sea level rise use case. C.T. participated in use case analysis and provided suggestions on ontologies to be used. X. M. prepared the figures. All authors contributed to the writing of the paper.

Tables

Table 1 | A list of ontologies and schemas, their namespace prefixes and corresponding URIs

Full name	**Namespace prefix**	**Namespace URI**
The PROV Ontology	prov	http://www.w3.org/ns/prov#
The Bibliographic Ontology	bibo	http://purl.org/ontology/bibo/
The Dublin Core Metadata Initiative Metadata Terms	dct	http://purl.org/dc/terms/
The GCIS Ontology	gcis	http://data.globalchange.gov/ontology
The Resource Description Framework Schema	rdfs	http://www.w3.org/2000/01/rdf-schema#
The eXtensible Markup Language Schema	xsd	http://www.w3.org/2001/XMLSchema#

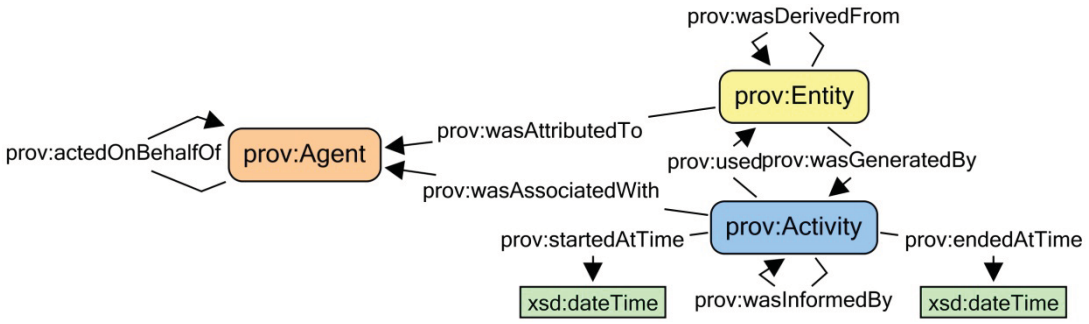
Figures



Metadata element	Record
Type	Bibliographic Resource / Figure
Identifier	2.26
Title	Global Sea Level Rise
Description	Estimated, observed and possible amounts of global sea level rise from 1800 to 2100. Proxy estimates (Kemp et al. 2012) (for example, based on sediment records) are shown in red (pink band shows uncertainty), tide gauge data in blue (Church and White 2011a), and satellite observations are shown in green (Nerem et al. 2010). The future scenarios range from 0.66 feet to 6.6 feet in 2100 (Parris et al. 2012). Higher or lower amounts of sea level rise are considered implausible, as represented by the gray shading. The orange line at right shows the currently projected range of sea level rise of 1 to 4 feet by 2100, which falls within the larger risk-based scenario range. The large projected range reflects uncertainty about how glaciers and ice sheets will react to the warming ocean, the warming atmosphere, and changing winds and currents. As seen in the observations, there are year-to-year variations in the trend. (Figure source: Josh Willis, NASA Jet Propulsion Laboratory)
Creator	Josh Willis, NASA Jet Propulsion Laboratory
Source	Chapter 2 in draft report of the third National Climate Assessment
Publisher	U.S. Global Change Research Program
Date	01/11/2013
References	Kemp, A.C., B.P. Horton, J.P. Donnelly, M.E. Mann, M. Vermeer, and S. Rahmstorf, 2012: Climate related sea-level variations over the past two millennia. *Proceedings of the National Academy of Sciences of the United States of America*, **108,** 11017-11022. doi: 10.1073/pnas.1015619108 Church, J.A. and N.J. White, 2011a: Sea-level rise from the late 19th to the early 21st century. *Surveys in Geophysics*, **32,** 585-602. doi: 10.1007/s10712-011-9119-1 Nerem, R.S., D.P. Chambers, C. Choe, and G.T. Mitchum, 2010: Estimating mean sea level change from the TOPEX and Jason altimeter missions. *Marine Geodesy*, **33,** 435-446. doi: 10.1080/01490419.2010.491031 Parris, A., P. Bromirski, V. Burkett, D. R. Cayan, M. Culver, J. Hall, R. Horton, K. Knuuti, R. Moss, J. Obeysekera, A. Sallenger, and J. Weiss, 2012: Global sea Level Rise Scenarios for the United States National Climate Assessment. NOAA Tech Memo OAR CPO-1. 37 pp. Accessible at: http://scenarios.globalchange.gov/sites/default/files/NOAA_SLR_r3_0.pdf

Figure 1 | Collected metadata records of a figure using Dublin Core metadata terms. The global sea level rise figure on the left is from the draft report of the third National Climate Assessment in the United States. In the table on the right, the metadata element 'References' is from the DCMI extended terms [7], and the others are from the core elements [6].

a



b

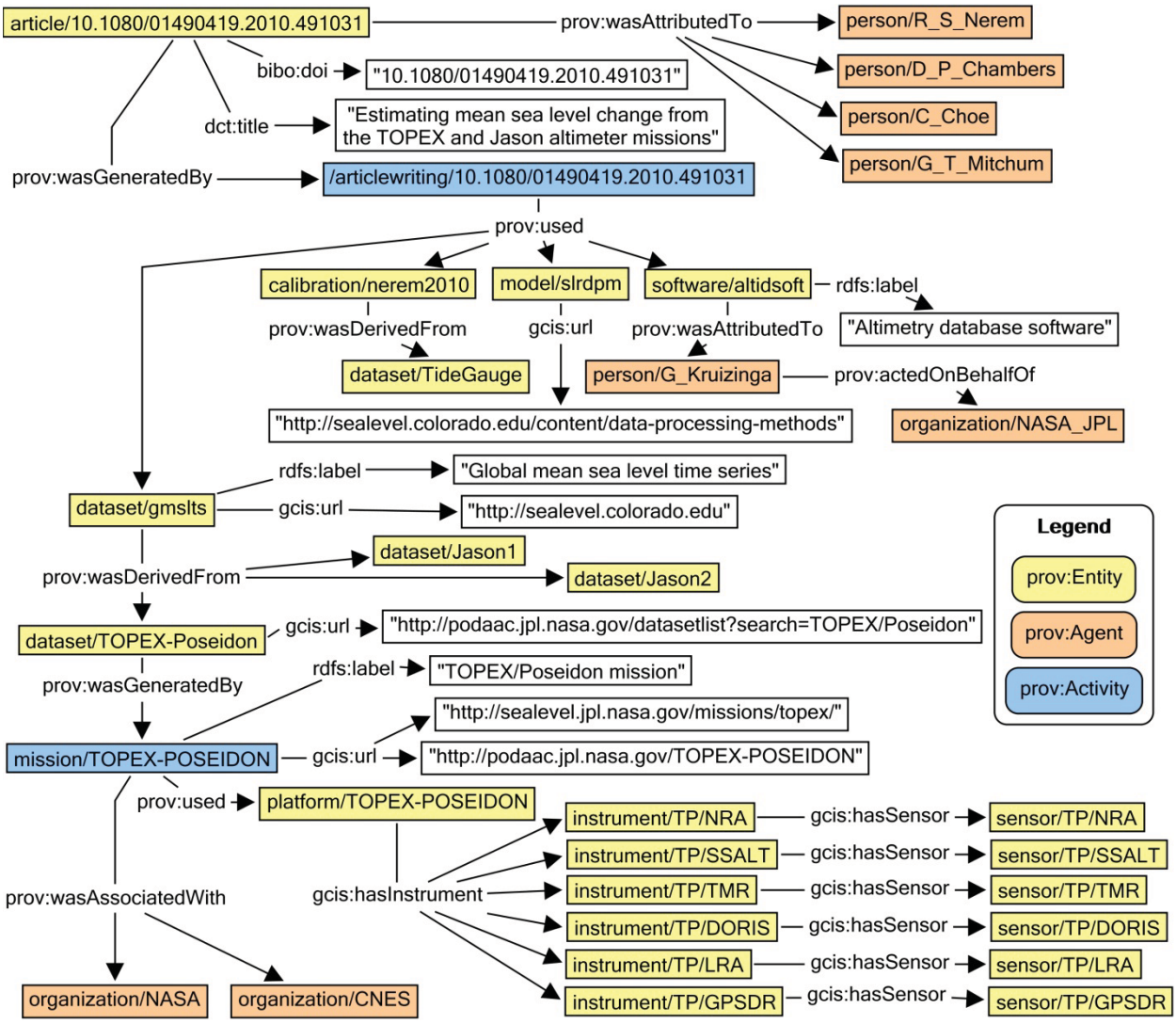


Figure 2 | A part of the provenance information of a journal article. **a**, Three starting-point classes/categories in the PROV-O and the properties that relate them. **b**, conceptual map of the

provenance information captured by using PROV-O and a few other ontologies. In (**a**) and (**b**) classes/categories are represented by rounded boxes, and instances are in rectangular boxes. The first term of the record in each colored rectangular box represents the class/category of the instance, and the color of that box represents the corresponding super class/category in the PROV-O. A colorless rectangular box represents a literal record or a Web address. In (**b**) the record in each colored rectangular box is a URI, and the common root address 'http://data.globalchange.gov/' is omitted from all records here to save space. The provenance information of 'dataset/TOPEX-Poseidon' is shown in detail and the information of 'dataset/Jason1' and 'dataset/Jason2' is omitted here. The namespace prefix of a class or property in (**a**) and (**b**) represents its source ontology, for which the details are listed in Table 1.