Provenance Capture in Data Access And Data Manipulation Software **[1]**

Submitted by pwest on Fri, 2014-01-03 15:12 **Event:** Winter Meeting 2014 [2]

Abstract:
There is increasing need to trace back the origins of data products, whether images or charts in a report, data obtained from a sensor on an instrument, a generated dataset referenced in a research paper, in government reports on the environment, or in a publication or poster presentation. Yet, most software applications that perform data access and manipulation keep only limited history of the data, i.e. the provenance. Imagine the following scenario: There is a figure in a report showing multiple graphs and plots related to global climate, the report is being drafted for a government agency. The graphs and plots are generated using an algorithm from an iPython Notebook, developed by a researcher who is using a particular data portal, where the algorithm pulls data from four data sets from that portal. That data is aggregated together over the time dimension, constrained to a few parameters, accessed using a particular piece of data access software, and converted from one datatype to another datatype; All the processing on the data sets was conducted by three different researchers from a public university, on a project funded by the same government agency requesting the report, with one Principal Investigator and two Co-Investigators. In this scenario, today we're lucky to get a blob of text under the figure that might say a couple things about the figure with a reference to a publication that was written a few years ago. Data citation, data publishing information, licensing information, and provenance are all lacking in the derived data products.

What we really want is to be able to trace the figure all the way back to the original datasets, including what was done to those datasets; and to see information about the researchers, the project, the agency funding, the award, and the organizations collaborating on the project. In this paper we discuss the need for such information and traceback features, as well as new technologies and standards that can help us become better data stewards. Specifically, we will talk about the new PROV recommendation from the W3C, recently published, and existing and new features in the OPeNDAP software stack that can help facilitate the incorporation of citation, licensing, and provenance information and the ability to click through to retrieve that information.

Collaboration Area: Semantic Web [3]
Attachments for download: ESIPWinter2014_OPeNDAP_PatrickWest.pptx [4]

Author(s):
Name: Patrick West [5]
Email: westp@rpi.edu [6]

Name: Peter Fox [7]
Organization(s): TWC [8]
Email: pfox@cs.rpi.edu [9]

Name: Deborah L. McGuinness [10]
Organization(s): Rensselaer Polytechnic Institute [11]
Email: dlm@cs.rpi.edu [12]

Name: James Gallagher [13]
Organization(s): OPeNDAP [14]

Name: Dan Halloway [15]
Organization(s): OPeNDAP [14]

Name: Nathan Potter [16]
Organization(s): OPeNDAP [14]

Keywords: OPeNDAP [17]
provenance [18]

Source URL: http://commons.esipfed.org/node/1960

Links:
[1] http://commons.esipfed.org/node/1960
[2] http://commons.esipfed.org/taxonomy/term/1029
[3] http://commons.esipfed.org/collaboration-area/semantic-web
[4] http://commons.esipfed.org/sites/default/files/ESIPWinter2014_OPeNDAP_PatrickWest.pptx
[5] http://commons.esipfed.org/node/1942
[6] mailto:westp@rpi.edu
[7] http://commons.esipfed.org/node/1356
[8] http://commons.esipfed.org/taxonomy/term/730
[9] mailto:pfox@cs.rpi.edu
[10] http://commons.esipfed.org/node/627
[11] http://commons.esipfed.org/taxonomy/term/221
[12] mailto:dlm@cs.rpi.edu
[13] http://commons.esipfed.org/node/412
[14] http://commons.esipfed.org/taxonomy/term/216
[15] http://commons.esipfed.org/node/1958
[16] http://commons.esipfed.org/node/1959
[17] http://commons.esipfed.org/taxonomy/term/318
[18] http://commons.esipfed.org/taxonomy/term/928