Integrating Provenance into an Operational Data Product Information System

Stephan Zednik, James Michaelis, and Peter Fox

Rensselaer Polytechnic Institute, Troy NY 12180, USA

Abstract. Knowledge of how a science data product has been generated is a critical component to determining its fitness-for-use for a given analysis. One objective of science information systems is to allow users to search for data products based on a wide range of criteria; spatial and temporal extent, observed parameter, research domain, and organizational project are common search criteria. Currently, science information systems are geared towards helping users find data, but not in helping users determine how the products were generated. An information system that exposes the provenance of available data products, that is what observations, assumptions, and science processing were involved in the generation of the data products, would contribute significant benefit to user fitness-for-use decision-making.

In this work we discuss semantics-driven provenance extensions to the Virtual Solar Terrestrial Observatory (VSTO) information system. The VSTO semantic web portal uses an ontology to provide a unified search and product retrieval interface to data in the fields of solar, solar-terrestrial, and space physics. We have developed an extension to the VSTO ontology that allows it to express item-level data product records. We will show how the Open Provenance Model (OPM) and the Proof Markup Language (PML) can be used to express the provenance of data product records. Additionally, we will discuss ways in which domain semantics can aid in the formulation - and answering - of provenance queries. Our extension to the VSTO ontology has also been integrated with a solar-terrestrial profile of the Observation and Measurement (O&M) model to support domain-specific descriptions of solar-terrestrial observations; we utilize this integration to connect observation events to the data product record lineage.

Our additions to the VSTO ontology will allow us to extend the VSTO web portal user interface with search criteria based on provenance and observation characteristics. More critically, provenance information will allow the VSTO portal to display important knowledge about selected data records; what science processes and assumptions were applied to generate the record, what observations the record derives from, and the results of quality processing that had been applied to the record and any records it derives from. We conclude by showing our interface for showing record provenance information and discuss how it aids users in determining fitness-for-use of the data.

Keywords: provenance, information systems, observation model, solar-terrestrial.