Using Domain Requirements to Achieve Science-Oriented Provenance

Eric Stephan[1], Todd Halter[1], Terence Critchlow[1], Paulo Pinheiro da Silva[2], and Leonardo Salayandia[2]

[1] Pacific Northwest National Laboratory, Richland WA, USA
[2] University of Texas at El Paso, El Paso TX, USA

Abstract. The US Department of Energy (DOE) Atmospheric Radiation Measurement Program (ARM) is adopting the use of formalized provenance to support observational data products produced by ARM operations and relied upon by researchers. Because of the diversity of needs in the climate community provenance will need to be conveyed in a domain-oriented context. This paper explores a use case where semantic abstract workflows (SAW) are employed as a means to filter, aggregate, and contextually describe the historical events responsible for the ARM data product the scientist is relying upon.

1 Introduction

What is the right level of provenance, disseminated to the right audience, in the right scientific context? This is a continual question facing the Department of Energy's Atmospheric Radiation Measurement (ARM) Program [1], especially as a diversity of audiences such as climate modelers, and researchers use the ARM data products in new and innovative ways. ARM is currently advancing the way day-to-day tasks relating to data capture, processing, and reprocessing, error detection, and troubleshooting in analytical methods by formally adding a provenance component to preserve the workflow history of tasks. This year the ARM Data Management Facility will manage the data flow for over 420 sensors located around the world, and will be ingesting one half terabyte of observations daily. As sensor data is collected, ARM transforms the ingested data into a uniform format, performs quality control, reprocessing, and transfers the finished products to the ARM Archive. From an operational standpoint it is foreseen that the number of ARM data products will continue to increase significantly, dwarfing today's complexity of algorithm interdependency. As demands on data increase the need for provenance is challenging our capability of properly supporting data and product attribution.

2 Issues with Scientific Outreach

For ARM sensor data stream processing and reprocessing, a standard Integrated Software Development Environment (ISDE) workflow has been adapted that is

D.L. McGuinness, J.R. Michaelis, and L. Moreau (Eds.): IPAW 2010, LNCS 6378, pp. 301–303, 2010.

comprised of all or a subset of the following steps: Initialize, Get, Translate, Scientific Analysis, Translate, and Put. While on the surface these steps seem trivial each step relies upon detailed algorithms for processing the data, and each algorithm must iterate through the stream to operate on each sample. Part of the provenance challenge is retaining knowledge about how the data was processed that meets the needs of both operational staff and downstream researchers.

Provenance at multiple tiers [2] is required to provide relevant information for operations and researchers. Each tier has a different focus and resolution. The first tier that represents the lowest resolution of provenance depicts lineage because ARM data products are highly interdependent. This information is not only invaluable to the researcher in terms of knowing what went into making a product, it is vital to addressing cascading errors produced when erroneous products are relied upon by downstream data processes. The underlying tiers depict provenance as a "hedge" or forest of ordered acyclic graphs. The hedge tier provides provenance and referential information common to all samples being processed at the component level within the ISDE workflow, and each acyclic graph within the context of the hedge represents the third tier that we refer to as the "branch". The current approach is to capture provenance for every step within the ISDE workflow, analyze the provenance from an operational standpoint, and retain a subset of provenance to be used by researchers. A challenge is that of only retaining provenance useful for scientific understanding as data products are archived for future dissemination to researchers.

3 Developing Domain-Oriented Provenance Requirements

One approach to retain provenance most useful for scientific means is to restrict its capture to knowledge with a well-defined usage. Because provenance needs will be diverse, a knowledge-driven strategy is suggested to identify provenance in support of ARM's diverse research community.

A common understanding of the ISDE workflow by domain experts and computer scientists is a requirement to understand where and how provenance needs to be permanently archived. For instance, many steps (and sub-steps) of the workflow are required to support the execution of the workflow, and from an operational standpoint need to verify datasets being pre-staged for translation. However, from a scientific perspective, this information needs to be conveyed in an aggregate perspective. In other cases, scientists need a high resolution perspective of the samples being translated, but need a filtered view of what actually occurred. In both cases, scientists need to be able to describe the ISDE workflow in terms of scientific steps, and of using these scientific steps to identify the ones that they need provenance.

Our strategy is to have different domain expert focus groups to describe their understanding of the ISDE workflow through the use of semantic abstract workflows (SAW) [3] and for the computer scientists to map the SAW steps (i.e., SAW methods) into concrete ISDE workflow tasks. One immediate benefit of

this approach is that the workflow would be described in scientific terms. Another benefit is that non-scientific steps are going to be abstracted away from the workflow. From the ISDE SAW, domain experts should be able to identify (and rank) the steps of the workflow that require provenance. From the SAW, it is possible to identify feasible provenance use cases and, by using the mappings between SAW methods and workflow tasks, to anticipate the content of the provenance for each task of the workflow. Moreover, the SAW can be use to determine provenance use cases that cannot be implemented because of unintentional uses of ARM data requiring knowledge not captured in the current provenance encoding.

References

1. Atmospheric Radiation Measurement Climate Research Facility: ARM Annual Report. U.S. Department of Energy, DOE/SC-ARM-0706 (2007)
2. Stephan, E.G., Halter, T.D., Ermold, B.D.: Leveraging The Open Provenance Model as a Multi-Tier Model for Global Climate Research. In: Proc. of 3rd International Provenance and Annotation Workshop, IPAW 2010 (2010)
3. Pinheiro da Silva, P., Salayandia, L., Gandara, A., Gates, A.Q.: CI-Miner: Semantically Enhancing Scientific Processes. Earth Science Informatics 2(4), 249–269 (2009)