Knowledge Provenance in Virtual Observatories: Application to Image Data Pipelines

Peter Fox
HAO/ESSL/NCAR
PO Box 3000
Boulder, CO 80307
(1) 303-497-1511
pfox@ucar.edu

Deborah McGuinness
Tetherless World Constellation
Rensselaer Polytechnic Institute
110 8th St, Troy NY 12180
(1) 518-276-4404
dlm@cs.rpi.edu

Paulo Pinheiro da Silva
Department of Computer Science
University of Texas El Paso
El Paso, Texas, 79968-0518
(1) 915 - 747-6827
paulo@utep.edu

ABSTRACT

Scientific data services are increasing in usage and scope, and with these increases comes growing need for access to provenance information. Our goal is to design and implement an extensible provenance solution that is deployed at the science data ingest time. In this paper, we describe our work in the setting of a particular set of data services in the area of solar coronal physics. The paper focuses on one existing federated data service and one proposed observatory. Our claim is both that the design and implementation are useful for the particular scientific image data services we designed for, but further that the design provides an operational specification for other scientific data applications. We highlight the need for and usage of semantic technologies and tools in our design and implemented service.

Categories and Subject Descriptors

H.2.5 {**Heterogeneous Databases**}: {Data translation}, I.2.4 [**Knowledge Representation Formalisms and Methods**]: Relation systems, Representation languages, Representations (procedural and rule-based), Semantic networks.

Keywords

Provenance; Image processing; semantics: markup, explanation, justification.

1. INTRODUCTION

Our goal is to create a next generation virtual observatory that includes an extensible representation for provenance for data ingest systems. Further, we consider provenance to be a first class item and our system will support semantically-enabled queries over the provenance as well as using provenance to filter data requests. In order to test our design, we are implementing our work using a domain in solar coronal physics. Our initial target is the Advanced Coronal Observing System (ACOS) currently operated at the Mauna Loa Solar Observatory (MLSO). The design is also expected to be implemented in the proposed CoSMO (Coronal Solar Magnetism Observatory).

We illustrate our setting in Figure 1 which is an abstracted representation of a typical data ingest pipeline for solar physics data streams. From the perspective of a provenance system, some aspects of the service are worth highlighting. First, data (represented in square boxes in the figure) passes through a number of stages and is potentially subject to a significant amount of possibly complex processing steps. Second, each collection and processing pass, including analysis and manipulations by humans, provides a place where provenance information could be and should be collected and represented. Third, quality control loops provide additional opportunities for provenance collection (and inspection).

The motivation for this project arose from our experiences designing and deploying a solar terrestrial physics virtual observatory system [1, 2] and from numerous discussions with the ‘data’ providers (i.e. roles) in Fig. 1. Among their remarks were the following:

- Data is being used in new ways and we frequently do not have sufficient information on what happened to the data along the processing stages to determine if it is suitable for a use we did not envision;
- We often fail to capture, represent and propagate manually generated information that need to go with the data flows;
- Each time we develop a new instrument, we develop a new data ingest procedure and collect different metadata and organize it differently. It is then hard to use with previous projects.

Further, when science data and information (often in the form of graphical images as is the case in our initial deployment)) are made available to an end-user (any of the roles in Fig. 1), it often happens after a number of data filtration and processing steps. As a consequence, any important metadata and/or documentation that may be needed to answer questions about the provenance may not have been generated, saved, propagated or be in a form or location that can be utilized (at all, or without significant effort or expertise). Virtual Observatories are particularly prone to this information gap. Thus, this project traces the entire pipeline and accounts for all roles, processes and metadata as they relate to use cases, which require provenance.

2. USE CASES

Use Case Development: After discussion and several meetings with the science project participants, we developed an initial set of use cases, which reflect a range of actual questions that are asked but at present cannot be answered in any routine or repeatable manner.

- Who (person or program) added the comments to the science data file for the best vignetted, rectangular polarization brightness image from January, 26, 2005 1849:09UT taken by the ACOS Mark IV polarimeter?
- What was the cloud cover and atmospheric seeing conditions during the local morning of January 26, 2005 at MLSO?
- Find all good images on March 21, 2008.
- Why are the quick look images from March 21, 2008, 1900UT missing?
- Why does this image look bad?

While we are considering all of these use cases in developing SPCDIS, we chose to begin with a particular set of images, i.e. quick look images and operator log integration.

Fig. 2 shows an initial implementation of provenance and inference applied to the use case: why are quick look images missing or are of poor quality. This corresponds to a specific portion of the data pipeline for an Imaging Photometer (CHIP) instrument. For details on the PML Source, Node Set and Inference Engine see [3]. In this diagram, the generation of the original quick look image (in GIF format) is combined with timestamp and observer log (which are ASCII text files) to create an extended quick look, in essence a marked up image (with PML) that can be displayed, indexed and searched, et c. The EQL App and Log Parser have been developed so far.

3. INFERENCE WEB (IW) AND THE PROOF MARKUP LANGUAGE (PML)

Inference Web [4] is a knowledge provenance infrastructure that supports comprehensive explanation capabilities. Those capabilities include interoperable explanations of sources (i.e. sources published on the Web or accessible from files), assumptions, learned information, and answers associated with inferred or stated conclusions. This knowledge provenance information may be used to improve users trust regarding those conclusions and thus may make systems including knowledge provenance support more actionable. Inference Web aims to make recommendations from intelligent systems more understandable, trustworthy, and usable by providing knowledge provenance infrastructure. Here we focus on explaining scientific information systems and knowledge provenance for scientific for scientific workflow and its data may include datasets, visualizations, and simulations – thus expanding the range of data types that the knowledge provenance infrastructure must support.

Inference Web provides the Proof Markup Language (PML) [4, 5], to encode justification information about basically any kind of response produced by agents. PML is an RDF based language defined by a rich ontology of provenance and justification concepts, which describe the various elements of automatically, generated proofs. PML may be used to include "standard" proofs of any sentence, thus encoding an entire rationale for a particular conclusion. This may be referred to as the proof and/or justification for the conclusion. Each PML component can reside in a uniquely identified document published on the Web separately from the others.

Probe-It! [6] is a browser suited to graphically rendering provenance information associated with results coming from both inference engines and scientific workflows. In this sense, Probe-It! does not actually generate content (i.e. logging or capturing provenance information); instead it is assumed that users will provide Probe-It! with end-points of existing provenance resources to be viewed. The task of presenting provenance in a useful manner is difficult in comparison to the task of collecting provenance. Here our main interest in terms of provenance visualization is on the use of Probe It! by scientists to better understand imperfection of CHIP images.

4. DISCUSSION AND CONCLUSIONS

To date, we have provided a valuable addition to one part of an image processing data pipeline for solar images taken by an imaging photometer at the Mauna Loa Solar Observatory in Hawaii. The creation of enhanced quick-look images in support of answering questions such as: "what were the weather and observing conditions for this quick look image and why does it look bad?" is providing significant added value to those monitoring the data pipeline and instrument performance. In the presented example, we have introduced structured observer log information into the explanation documentation, which was not previously available. We have also created PML instances and used two sets of tools to browse and search these explanations.

The next stage in this work will involve moving to other parts of the pipeline in Fig.1, i.e. earlier in the pipeline benefits instrument designers and project scientists, while later in the pipeline benefits data analysts and end-users. We also plan to complete the documentation of the remaining instrument data pipelines at MLSO and convert them to workflow-based systems.

5. ACKNOWLEDGMENTS

The Semantic Provenance Capture in Data Ingest Systems (SPCDIS) project is funded by the National Science Foundation, Office of Cyber Infrastructure under the SDCI program, grant number OCI-0721943. The National Center for Atmospheric Research is operated by the University Corporation for Atmospheric Research with substantial sponsorship from the National Science Foundation. We also acknowledge support from NSF CyberShARE - HRD-0734825 to UTEP.

REFERENCES

[1] Deborah McGuinness, Peter Fox, Luca Cinquini, Patrick West, Jose Garcia, James L. Benedict, and Don Middleton. The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research. In the Proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07). Vancouver, British Columbia, Canada, July 22-26, 2007.

[2] Peter Fox, Deborah L. McGuinness, Luca Cinquini, Patrick West, Jose Garcia, James L. Benedict, and Don Middleton. Ontology-supported Scientific Data Frameworks: The Virtual Solar-Terrestrial Observatory Experience. To appear in Computers and Geosciences Journal.

[3] Deborah L. McGuinness, Li Ding, Paulo Pinheiro da Silva, Cynthia Chang - PML 2: A Modular Explanation Interlingua. In ExaCt pp. 49-55, 2007. Also Stanford KSL Tech Report KSL-07-07.

[4] McGuinness, D. and Pinheiro da Silva, P. 2004, Explaining Answers from the Semantic Web: The Inference Web Approach. Web Semantics: Science, Services and Agents on the World Wide Web Special issue: International Semantic Web Conference 2003 - Edited by K.Sycara and J. Mylopoulous. 1(4). Fall, 2004. Also, Stanford KSL Tech Report KSL-04-03.

[5] Pinheiro da Silva, P., McGuinness, D., and Fikes, R. 2006, A Proof Markup Language for Semantic Web Services. Information Systems, 31(4-5), June-July 2006, pp 381-395. Prev. version, KSL Tech Report KSL-04-01.

[6] Del Rio, N. and Pinheiro da Silva. P., Probe-it! Visualization support for provenance. In Proceedings of the Second International Symposium on Visual Computing (ISVC 2), Lake Tahoe, NV, USA. Volume 4842 of LNCS, pages 732-741. Springer, 2007.