Semantic Provenance for Science Data Products: Application to Image Data Processing

Stephan Zednik, Peter Fox, Deborah L. McGuinness, Paulo Pinheiro da Silva, Cynthia Chang

Abstract—A challenge in providing scientific data services to a broad user base is to also provide the metadata services and tools the user base needs to correctly interpret and trust the provided data. Provenance metadata is especially vital to establishing trust, giving the user information on the conditions under which the data originated and any processing that was applied to generate the data product provided.

In this paper, we describe our work on a federated set of data services in the area of solar coronal physics. These data services provide a particular challenge because there is decades of existing data whose provenance we will have to reconstruct, and because the quality of the final data product is highly sensitive to data capture conditions, information which is not currently propagated with the data.

We describe our use of semantic technologies for encoding provenance and domain knowledge and show how provenance and domain ontologies can be used together to satisfy complex use cases. We show our progress on provenance search and visualization tools and highlight the need for semantics in the user tools. Finally, we describe how our methods are applicable to generic data processing systems.

1 Introduction

We aim to create a next-generation virtual observatory[1] with extensive provenance support. Provenance is a first-class concept in our system; with full support in our search, explanation, and visualization tools. We require a general provenance model that is applicable in a wide array of domains, and integrable with domain models, so that domain concepts can be modeled along with provenance concepts. For these reasons we have chosen to use the Proof Markup Language (PML) [3], [7] family of OWL ontologies as our provenance model. We show how PML can be used to model provenance causality chains, introduce our domain model, and show how the PML provenance model and our science domain models can be integrated in a manner that provides a rich provenance infrastructure, able to model complex scientific provenance relations.

We have chosen to test our system in the domain of solar coronal physics, using the Advanced Coronal Observing System (ACOS) as a testbed. The ACOS data products are the result of several data ingest pipelines, processing observations from three imaging instruments located at the Mauna Loa Solar Observatory (MLSO). The ACOS data pipelines are distributed data pipelines, operated in part at MLSO in Hawaii and the National Center for Atmospheric Research High Altitude Observatory (NCAR/HAO) in Boulder, CO. ACOS has been operational for over a decade, and has produced terabytes of data.

ACOS was chosen because its data pipelines are typical of data ingest systems and the vast quantity of existing data ACOS has generated in its decades of operation provides the opportunity to design a system geared to reconstruct, as well as capture, provenance. One of the ACOS data pipelines, the Chromospheric Helium-I Imaging Photometer (CHIP) Intensity Image pipeline, is illustrated in Figure 1. This high-level diagram is designed to show not just the process/artifact flow of the pipeline, but domain concepts that could

- *Stephan Zednik, Peter Fox, Deborah L. McGuinness, and Cynthia Chang are with the Rensselaer Polytechnic Institute, Tetherless World Constellation.*
- *Paulo Pinheiro da Silva is with the University of Texas at El Paso, Department of Computer Science.*

1. A virtual observatory is a collection of interoperating data archives and software tools which utilize the internet to form a scientific research environment in which research programs can be conducted.

and should be captured and represented in the provenance. In the pipeline, data (square boxes) passes through a number of stages (ovals) each of which can contain a number of complex processing, analysis, human interaction, and decision steps. Each of these stages contains domain-specific information (dotted-lined boxes) related to the data product provenance.

Of particular interest is information in the pipeline that is not a direct or inferred result of the data capture event. The Observer Log is a human-generated account of weather conditions and system status during the instrument observing schedule for the day. Bad weather conditions or instrument instability, noted in the log, can have significant negative effect on the quality of the data observations. This information is currently not propagated in the data pipeline nor do the data products reference it in any way, but is an invaluable reference in determining why an image has been given a low-grade quality assessment. This information is an important component of the origin of the data image and should be represented in its provenance.

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 Figure 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 needs to go with the data flows.

Further, when science data and visual representations of the data (such as the CHIP Quick Look images) are made available to the end-user, the product has often gone through a number of data filtration and processing steps. If thorough provenance metadata and processing documentation is not captured, propagated, and made available to the end-user; the data system is in effect a 'black box', and the end-user must blindly trust the science quality of the data product and long-term consistency of the pipeline processing.

Virtual Observatories are particularly prone to this information gap. This project traces the entire pipeline and accounts for all roles, processes, and metadata as they relate to use cases which require provenance.

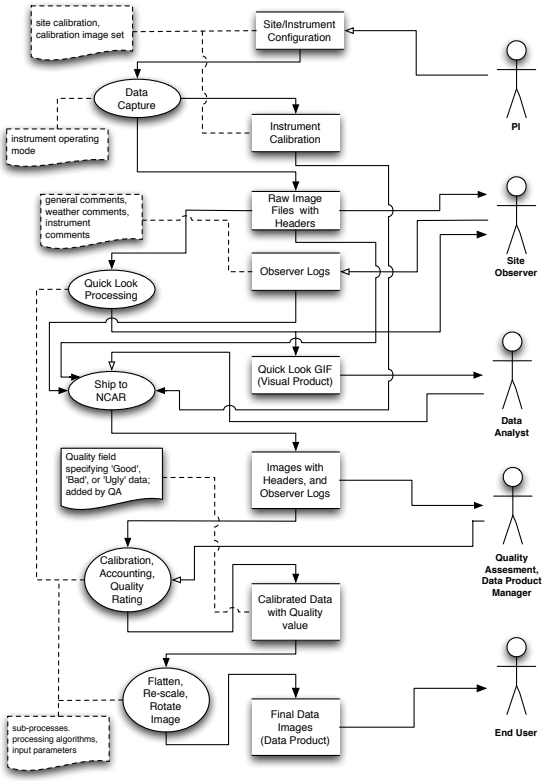

Fig. 1.
Chromospheric Helium-I Imaging Photometer (CHIP) Intensity Image pipeline

2 Use Cases

During discussions with science project participants, we developed an initial set of use cases which reflect real user questions that cannot presently be answered in any routine or automated manner:

- What were the cloud cover and seeing conditions during the observation period of *this* data product artifact?
- What calibrations have been applied to *this* data product artifact?
- Who (person or program) added the comments to the science data file for the best vignetted rectangular polarization brightness image from January 26, 2005 18:49:09UT taken by the ACOS Mark IV polarimeter?

- Find all *good* CHIP He 1083 nm intensity images on March 21, 2008.
- Why does *this* data look *bad*?

These use case scenarios mix domain and provenance terms in a manner that would make the question difficult to answer if the provenance and domain models are independent. The cloud cover and seeing conditions during observation periods is recorded, but it is never directly associated with nor propagated with processed data. We could adjust the data processing pipeline to pull this information from the observation records and propagate it with the data as metadata; but this would be a very heavy-handed approach to take for any and all information associated with a data product's generation and processing that a user may want to see. Instead, we intend to build the link in the provenance representation of the data product, so that by following the data product's causality graph the system can find the weather condition records, calibrations applied, quality control information, etc. that are associated with the data product's generation or processing.

These use case scenarios are representative of many different question types that are routinely asked for all data products produced by the ACOS data ingest pipeline, and we believe these are representative of use cases that are common in any science data pipeline application.

3 Provenance Requirements

We require a provenance infrastructure that supports queries, filtering, and reasoning by domain concepts. A design requiring the hardwiring of domain concepts into the provenance model, or into the system logic that accesses the provenance store, is undesirable because it will be difficult to maintain and extend, and furthermore, such a hardwired application will make interoperation more challenging.

The provenance infrastructure must also support existing data systems, requiring little to no modification of the processing pipeline; ACOS is a production system, and we do not have the opportunity to re-engineer it. The system should also support generating some amount of provenance for existing processed data. It is not feasible to reprocess all existing data, and doing so with the current pipeline may introduce discrepancies between the newly processed products and archived products processed on a earlier and different version of the pipeline. The provenance capture should be configurable such that as much provenance as possible can be generated based on our understanding of an earlier version of the pipeline, without forcing us to re-run data processing.

Finally, since ACOS is a distributed system the provenance infrastructure must also work as a distributed system. Provenance should be gathered where processing occurs and made available as part of a distributed provenance store.

4 Data Model

4.1 Provenance Representation

To support our provenance requirements we have elected to use OWL ontologies to model both domain and provenance concepts. The provenance and domain base ontologies are independent, but the system's individuals reference both models (via multiple-inheritance), so queries, filtering and reasoning by either domain or provenance concepts are supported. This design supports our desire to build a maintainable system that refrains from hardcoding solar terrestrial concepts into the base provenance model or provenance logic.

We have chosen as our provenance model the Inference Web [6] Framework's Proof Markup Language [3], [7] (PML) because of its capabilities in representing conclusions, justifications (inference and source usage), and explanations. Another particularly useful aspect of the PML model is its separation of the process engine and process rule concepts. By defining these concepts separately, PML can represent both the process that was executed (PML InferenceEngine) and the rule (PML InferenceRule) by which the executed process operated. Another way to view this concept separation is that PML can capture both execution history and execution purpose. Inference rules are a pivotal concept of the justification aspect of PML and provide a clear mechanism for relating domain concepts to a provenance causality graph. While other provenance models such as Open Provenance Model (OPM) could have provided some of the foundation provided in PML, we found some of the core representational constructs such as those mentioned above to be well suited for our applications. For further analysis of the relative benefits of PML and OPM, see 'Towards Usable and Interoperable Workflow Provenance: Emperical Case Studies using PML' [4] and 'Domain Knowledge and Provenance in Science Data Systems' [5].

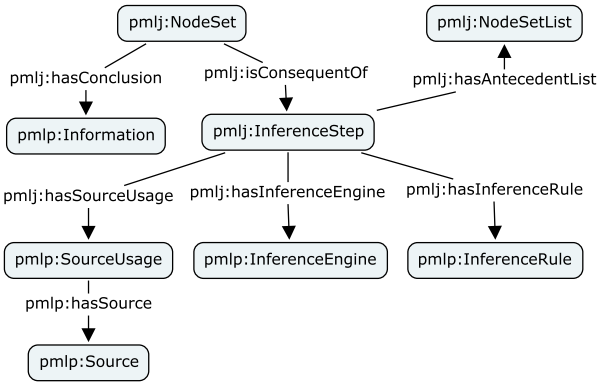

Fig. 2.
Basic PML NodeSet

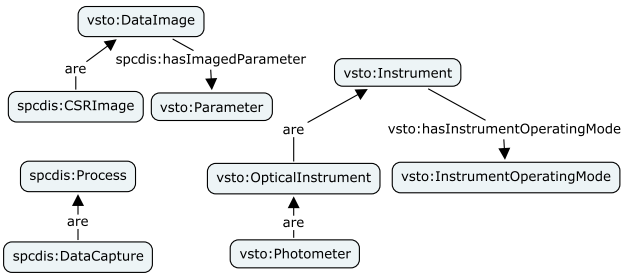

Fig. 3.
VSTO (vsto) and Science Data Processing (spcdis) domain concepts

In PML, a piece of information (the conclusion) and its justification(s) are modeled as a nodeset. A nodeset justification, known as an inference step, is used to describe the engine or source and the rule used to generate the nodeset conclusion. Each inference step may specify a list of nodesets, known as antecedents, whose conclusions it is dependent upon. The antecedent relations between nodesets can be used to build a causality graph, or explanation, for the conclusion of the nodeset. The fundamental classes and properties of a PML nodeset are shown in Figure 2.

4.2 Domain Representation

The VSTO[2] solar-terrestrial ontology, developed during our previous experience deploying a semantic virtual observatory [1], [2] will be used as one of our core science domain models. The VSTO ontology provides a model for data products, instruments, and parameters related to solar-terrestrial data systems. The VSTO ontology does not currently describe the processing that occurs in a typical science data ingest pipeline (calibrations, transformations, data filtering, quality control processes, etc.) so we are developing our own science data processing ontology based on experience gained during this[3] project and similar work with the MDSA[4] project.

Figure 3 illustrates some domain model concepts from the VSTO and (in-development) science data processing ontologies that we will be integrating with the ACOS provenance. Of particular interest in this example is the vsto:InstrumentOperatingMode concept, which is defined as the *configuration and process that allows an instrument to produce the required signal*. In the solar terrestrial domain terminology, an operating mode is not treated as an input configuration to a process, not as an artifact, but as a description of the state of the instrument and the entailing process for data capture. It is a description of how an instrument performs data capture of a specific parameter type. In fact, vsto:InstrumentOperatingMode was originally modeled as a subclass of the class vsto:AbstractProcess. The VSTO InstrumentOperatingMode and science data processing ontology DataCapture concepts relate to each other in much the same way the PML InferenceRule and InferenceEngine relate, and in the next section we will show how they can be integrated.

4.3 Provenance and Domain Model Integration

The provenance and domain ontology concepts are integrated not in the model definitions, but in the individuals' declarations by taking advantage of OWLs natural support for multiple-inheritance. Where it is deemed beneficial to express both domain and provenance concepts, individuals (ontology class instances) are defined with multiple types, one type from the provenance model and at least one type from the domain ontologies. As an example, the science data processing ontology may define an individual spcdis:FlatFieldCalibration of type spcdis:Calibration and type pmlp:InferenceRule. The use case *'What calibrations have been applied to*

2. Virtual Solar Terrestrial Observatory, http://vsto.org/
3. Semantic Provenance Capture in Data Ingest Systems
4. Multi-Sensor Data Synergy Advisor, http://tw.rpi.edu/portal/MDSA

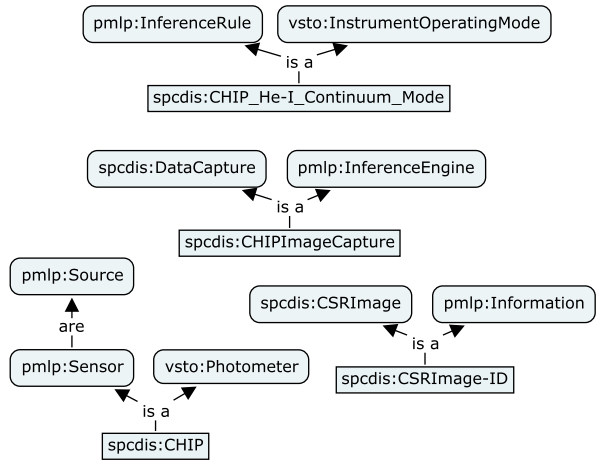

Fig. 4.
Individuals integrating provenance and domain models

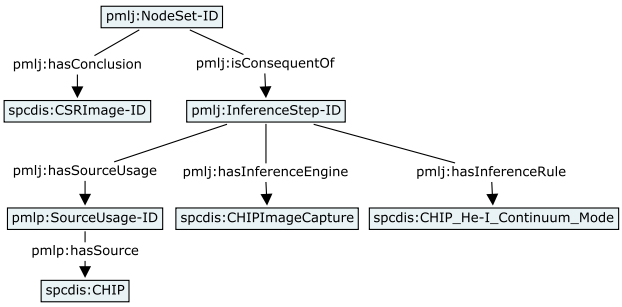

Fig. 5.
PML NodeSet comprised of domain model integrated individuals

this data product artifact?' may now be answered by querying the data product's causality graph for inference rule's of type spcdis:Calibration used by any nodeset's justification.

Figure 4 shows how individuals of the domain concepts illustrated in Figure 3 may be mapped to PML provenance concepts. The instrument operating mode instance is modeled not as the conclusion of a some nodeset's justification that acts as an antecedent to the data capture nodeset's justification, but as the rule by which data capture occurs. The individual spcdis:CHIP is defined as both a vsto:Photometer and a pmlp:Sensor, allowing it to be both the source of the conclusion of the data capture justification as well as assert any properties for which vsto:Photometer is in the domain.

We can now model a PML NodeSet using individuals that have type and properties from domain ontologies, as shown in Figure 5. Integrating domain types and properties with the provenance-based causality graph allows us to answer complex use case scenarios such as *'What calibrations have been applied to this data product artifact?'* by performing reasoning on domain concepts integrated with the individuals in the provenance.

5 Provenance Capture

To assist in PML generation, we describe a type of program referred to as a PML data annotator. A PML data annotator is a simple program whose sole purpose is to capture the provenance of a single decision/process in a decision system and encode that provenance as a PML nodeset. PML data annotator programs are run as components of a workflow; either as part or separate to the actual decision processing. When run as part of the data processing, the PML data annotator invokes the inference engine directly; extracting required processing inputs from antecedent nodesets and passing this information during inference engine invocation.

For the ACOS provenance capture we will utilize a parallel workflow, where PML data annotators do not directly invoke inference engines but reconstruct the processing of the existing data ingest pipelines. This architecture also supports our need for provenance generations for archived or pre-existing data products without preprocessing of the data. The PML data annotators are in in a workflow that simulates the processing of the data pipeline using analysis of existing artifacts and information about the data processing encoded in the PML data annotator configuration to reconstruct provenance. The PML data annotator workflow can be reconfigured to simulate different variations of the data processing pipeline to generate provenance from data processing pipelines that are no longer active.

6 Provenance Search

We will utilize the search and explanation capabilities of the Inference Web toolset to provide both a free text and guided search on provenance and domain concepts. Guided searches generate a SPARQL query on the provenance + domain RDF and free text searches currently performs a standard full text index search on the same. Search results

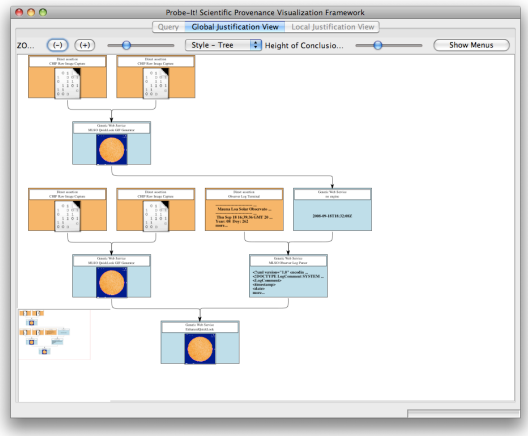

Fig. 6.
Probe-It visualizing the provenance of a CHIP QuickLook Visual Product

can be viewed using a number of tools including the Inference Web browser, where the user can explore the provenance encoding in detail, or in Probe-It!, a second generation PML provenance visualization tool using applet technology for greater graph functionality, introduced in the next section.

7 Provenance Visualization

Probe-It! [8] is graphical browser of PML-based provenance, developed by Cyber-ShARE at the University of Texas at El Paso. Probe-It! generates a causality graph for all antecedents to a specified nodeset and generates a visual representation (where applicable) for the conclusion of all nodesets in the graph. A Probe-It! visualization of the CHIP QuickLook[5] Visual Product is shown in Figure 6. Our primary interest in Probe-It for the ACOS provenance is in enabling scientists to better understand imperfections in and processing consequences upon science data images.

8 Discussion and Conclusion

To date we have reconstructed provenance for the Quick Look visual product from the CHIP Intensity data ingest and shown how weather condition information, previously not propagated, can provide value added to the data product. We have introduced our work constructing a domain-aware provenance store built using a generic, extensible provenance model and a solar-terrestrial domain model.

We described how this provenance store can be used to represent the relationships expressed in our use cases, and why these use cases are important in increasing user trust in the data products. While we did this in one workflow setting, since the use cases are representative of those in many scientific workflow settings, we believe this work provides a foundation for scientific workflow provenance applications. We have described in brief how we intend to use a parallel workflow to reconstruct and generate provenance, and the semantically-enabled provenance search, explanation, and visualization tools we will provide for the end users.

The next stage of our work will involve further modeling of data pipeline concepts in the ACOS provenance ontology, further documentation of the ACOS data pipelines, and construction of PML wrappers for newly documented sections of the data pipelines. We also intend to prototype semantic provenance faceted-search interfaces, move our free text search to the Apache Lucene text search engine, and develop new visual representations of nodeset conclusions in our visual provenance browser. Following the completion and testing of the ACOS application, we will separate our extensions that are ACOS-specific from those that are general to science applications and release scientific provenance module extensions for VSTO and PML ontologies and related wrapper support tools.

5. QuickLook images are lightly-calibrated visual approximations of the data image, generated in real-time and used primarily as quality checks on instrument operation

Acknowledgments

The SPCDIS[6] project is funded by NSF, Office of Cyber Infrastructure under the SDCI program, grant number OCI-0721943, and in collaboration with UTEP CyberShare Center, funded by NSF under the CREST program, grant number HRD-0734825.

References

[1] McGuinness, D., Fox, P., Cinquini, L., West, P., Garcia, J., Benedict, J., Middleton, D.: The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research. In the proceedings of the 19th Conference on Innovative Applications of Artificial Intelligence (IAAI). Vancouver, BC, Canada, July 2007, 1730-1737 and AI magazine, 29, #1, 65-76.

6. Semantic Provenance Capture in Data Ingest Systems

[2] Fox, P., McGuinness, D., Cinquini, L., West, P., Garcia, J., Benedict, J., Middleton. D.: Ontology-supported Scientific Data Frameworks: The Virtual Solar-Terrestrial Observatory Experience. Computers and Geosciences. Vol. 35, Issue 4, pp 724-738.

[3] McGuinness, D., Ding, L., Pinheiro da Silva, P., Chang, C.: PML 2: A Modular Explanation Interlingua. In ExaCt pp. 49-55 Also Stanford KSL Tech Report KSL-07-07 (2007)

[4] Michaelis, J., Ding, L., Shangguan, Z., Zednik, S., Huang, R., Pinheiro da Silva, P., Del Rio, N., McGuinness, D.: Towards Usable and Interoperable Workflow Provenance: Empirical Case Studies using PML. To appear in the Proceedings of the First International Workshop on the role of Semantic Web in Provenance Management, Chantilly, VA. (2009)

[5] Zednik, S., Fox, P., McGuinness, D.: Domain Knowledge and Provenance in Science Data Systems. To appear in Emerging Issues in e-Science: Collaboration, Provenance, and the Ethics of Data (IN13), AGU Fall 2009 Meeting, San Francisco, CA. (2009)

[6] McGuinness, D., Ding, L., Pinheiro da Silva, P.: 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.

[7] Pinheiro da Silva, P., McGuinness, D., Fikes, R.: 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 (June 2006)

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