

Semantic Provenance Capture in Data Ingest Systems <http://spcdis.hao.ucar.edu>

Project Background

The High Altitude Observatory at NCAR collects and serves many data streams for internal and community use within the solar and solar-terrestrial communities.

The goal of this project is to provide an extensible representation for provenance for data ingest systems. We focus on a set of solar coronal physics instruments, but with a view of the broader area of solar and solar-terrestrial, and terrestrial physics - the ACOS (Advanced Coronal Observing System) currently operated at the Mauna Loa Solar Observatory (MLSO) in Hawaii. The provenance work includes domain-independent portions geared for any data ingest system as well as a domain-literate module aimed at solar and solar-terrestrial physics.

Fig. 1 is a graphic representation of a typical data ingest pipeline for solar physics data streams. The data, in square boxes, passes through a number of stages and is subject to the use of processing, in the circles/ellipses, addition of metadata and influence by various human roles in the form of quality control loops. The levels of processing are numbered, 0, 1, 2, etc. QC stands for quality control. PI is the principal investigator. The motivation for this project arose from numerous discussions with the 'data' providers (i.e. roles) in Fig. 1.

When science data and information (often in the form of graphical images) are made available to an end-user (any of the roles in Fig. 1), it is after Level 3. 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 from Level 0 to 4 and accounts for all roles, processes and metadata as they relate to use cases, which require provenance.

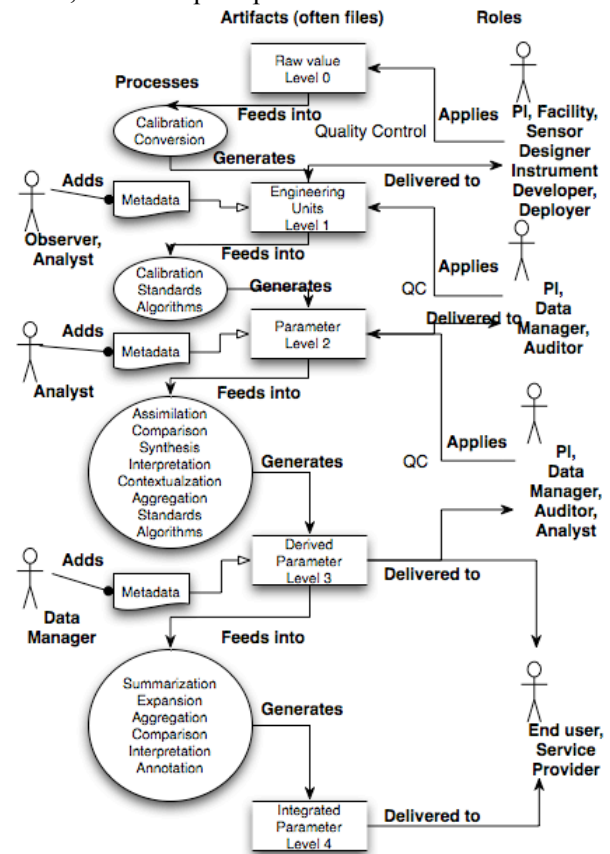


Figure 1: Generic representation of an instrument-based data ingest pipeline.

Use Case Development

Example use cases driving development:

- Who (person or program) added the comments to the science data file for the best vignettted, 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?**

Architecture schematic

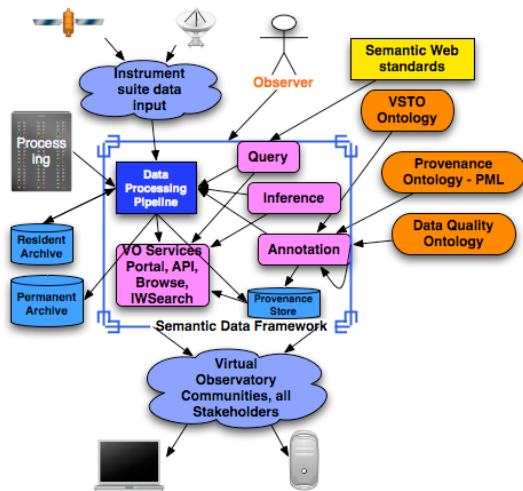


Figure 2: The SPCDIS architecture

Fig. 2 shows a schematic of the evolving SPCDIS architecture, which builds on the VSTO architecture. Added elements are: the contributions of the observer, PML and the Data Quality ontology, Annotation and the Provenance store. The annotation function is one element that has been the focus of our work to date, along with the use of browse (using Probe-It, not shown; see the demo at <http://iw.cs.utep.edu:8080/startprobeit/applet>) and search (Fig. 4) for the generated provenance.

Fig. 3 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 the top/ beginning of the data ingest documented in Fig. 3 for the CHIP instrument.

For details on the PML Source, Node Set and Inference Engine see <http://iw.rpi.edu>. 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. We have developed the EQL App and Log Parser. Next steps for this work involve: application to science images and engineering QC parameters.

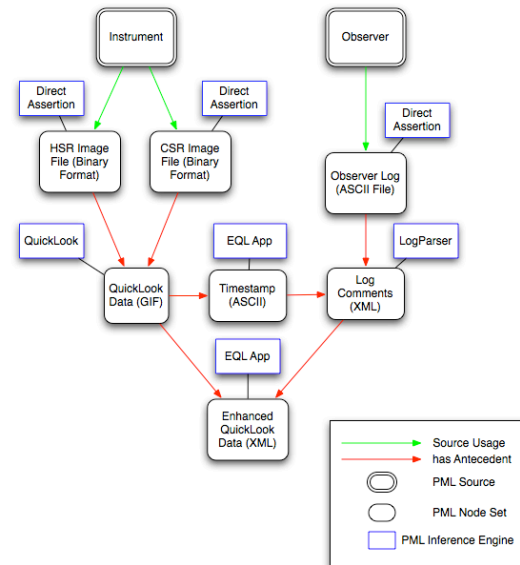


Figure 3: Provenance and inference implementation for the first SPCDIS demonstration. EQL is Enhanced Quick Look.



SPCDIS Search

Query:

Results 1 to 10 of 60 for 'quicklook from March 3 0:00 to March 3 23:59'

label	type	source	Date & Time	browse
08_05_03_10_00_00.eql	EQL	http://...	March 3, 2008, 10:00	image
08_05_03_10_15_00.eql	EQL	http://...	March 3, 2008, 10:15	image
08_05_03_10_30_00.eql	EQL	http://...	March 3, 2008, 10:30	image
08_05_03_10_45_00.eql	EQL	http://...	March 3, 2008, 10:45	image
08_05_03_11_00_00.eql	EQL	http://...	March 3, 2008, 11:00	image
08_05_03_11_15_00.eql	EQL	http://...	March 3, 2008, 11:15	image
08_05_03_11_30_00.eql	EQL	http://...	March 3, 2008, 11:30	image
08_05_03_11_45_00.eql	EQL	http://...	March 3, 2008, 11:45	image
08_05_03_12_00_00.eql	EQL	http://...	March 3, 2008, 12:00	image
08_05_03_12_15_00.eql	EQL	http://...	March 3, 2008, 12:00	image

1 2 3 4 5 6

Figure 4: Screen shot of SPCDIS structured search using time range reasoning. This interface is based on the IWSearch infrastructure.

Peter Fox, Deborah McGuinness, Paulo Pinheiro da Silva, Stephan Zednik, Jose Garcia, Li Ding, Nicholas Del Rio, Cynthia Chang, Thomas Zurbuchen