Fact: Scientific data services are increasing in usage and scope, and with these increases comes growing need for access to provenance information.

Science Project Goal: design and implement an extensible provenance solution that is deployed at the science data ingest time.

Provenance Infrastructure Goal: to design a reusable, interoperable provenance infrastructure.

Outcome: implemented provenance solution in one science setting AND operational specification for other scientific data applications.

Example Use Cases
1. What was the cloud cover and atmospheric seeing conditions during the local morning of January 26, 2005 at MLSO?
2. Find all good images on March 21, 2008.
3. Why are the Quicklook images from March 21, 2008, 1900UT missing?
4. Why does this image look bad?

From the Data Providers
Data is coming faster, in greater volumes and outstripping our ability to perform adequate quality control. 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.

Figure 1: Documentation of the Coronal Helium Imaging Photometer (CHIP) data processing pipeline, including the identification of specific artifacts, processes and people involved.

Figure 2: Provenance workflow for initial Quicklook deployment.

Building blocks
- The proof markup language (PML) provides an interlingua for capturing the information agents need to understand results and to justify why they should believe the results.
- The Inference Web toolkit provides a suite of tools for manipulating, presenting, summarizing, analyzing, and searching PML in efforts to provide a set of tools that will let end users understand information and its derivation, thereby facilitating trust in and reuse of information.

Benefits/Next Steps:
- Use of declarative provenance interlingua (PML).
- Supports integration of automatic and human-generated content into machine-driven processes/workflows.
- Multiple presentation views of provenance built using same infrastructure.
- Implementation of structured presentation of results use case.

Next: application to science images and engineering quality control use cases, expansion of search capabilities.

Figure 3: Probe-It! Screen shot displaying the main screen – full provenance trace is in the lower left corner (inset) and the main screen shows the zoomed view of the lower portion (see zoom rectangle in lower left inset) and provenance viewer (right) with formatted text description of items extracted from the observer log.

Additional Team members: Leo Salayandia, Aida Gandara, Jiao Tao, Honglei Zeng

Partially supported by NSF/OCI award OCI-0721943 and NSF CyberShARE - HRD-0734825.