Semantic Provenance Capture in Data Ingest Systems
Annual report and next year plans

**Project Background**

The project is entitled: SDCI: Data/NMI New/Improvement - Semantic Provenance Capture in Data Ingest Systems (SPCDIS) and originated in 2007 at the National Center for Atmospheric Research Earth and Sun Systems Lab, High Altitude Observatory (NCAR/ESSL/HAO), involving the University of Michigan, McGuinness Associates and the Trust group from the CyberSHARE project at the University of Texas at El Paso. In late 2008, the PI moved to the Tetherless World Constellation at RPI. We have a strong collaboration with The University of Texas at El Paso as the project has continued to develop this year. This report covers work through Aug. 31, 2011.

As a basis for the primary scientific data pipeline, 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 overall goal of this project is to provide an extensible representation for knowledge provenance for data ingest systems and demonstrate the robust nature of the provenance languages and tools so that this work can lead to substantial progress in related fields. As such, the provenance work includes domain-independent portions that are being shown to be applicable for general scientific data pipelines as well as domain-specific portions for the current implementation in solar and solar-terrestrial physics.

**Executive Summary of Progress**

The additional use cases developed in the previous year have been implemented (details of progress can found in the Working Groups and Use Cases sections of the project website at http://tw.rpi.edu/web/project/SPCDIS). The previous year’s activities, which focused on computer science areas: management of the provenance artifacts, especially their presentation at ontologies combined with domain vocabularies, presentation of the provenance (and related) results to end users. This year we shifted back to the domain needs of the HAO researchers, increasing research collaboration visits and beginning work on the final instrument that we originally proposed to implement the provenance pipeline for. The NCAR High Altitude Observatory, The University of Hawaii, and the University of Michigan, is building a COral Solar Magnetism Observatory (COSMO) facility for use by the solar physics research community (http://www.cosmo.ucar.edu/).

A key supporting instrument is the Coronal Multi-channel Polarimeter (CoMP). The CoMP instrument can observe the coronal magnetic field with a full FOV in the low corona (~1.03 to 1.5 Rsun), as well as obtain information about the plasma density and motion. Like Solar-C, CoMP records the intensity and the linear and circular polarization (Stokes I,Q,U,V) of the forbidden lines of Fe XIII at 1074.7 nm and also at 1079.8 nm. In addition to detecting the POS field direction and the LOS field strength, CoMP also
measures the LOS plasma velocity from Doppler observations in the wings of the line intensity (Stokes I), and the POS density from the ratio of the lines at 1074.7 and 1079.8 nm. The CoMP instrument was integrated into the Coronal One Shot coronagraph mounted on the equatorial spar at the Hilltop facility at the Sacramento Peak Observatory of the National Solar Observatory in 2010 and has begun acquiring routine scientific quality images as of October 1, 2010. Data has become available via the [Mauna Loa Solar Observatory](http://www.mlo.obs.nascom.nasa.gov/) web site as of May 1, 2011.

As we enter the latter stages of the project we have been actively incorporating provenance into the Semantic eScience Framework (SESF, NSF/OCI/STCI funded which has just completed its second year - [http://tw.rpi.edu/web/proj/S/SESF](http://tw.rpi.edu/web/proj/S/SESF)) via model and ontology evaluations and developments. This work has allowed to integrate the draft international standard (DISO-19156) for Observations and Measurements (O&M 2.0, Open Geospatial Consortium standard) into the model development, allowing us to evolve the original VSTO ontologies into a core SESF ontology and a Solar-Terrestrial Observations and Measurement (STOM) ontology to replace VSTO. The STOM and DQ ontologies can be found in the RPI eScience SVN at [https://scm.escience.rpi.edu/svn/public/ontologies/STOM/trunk/](https://scm.escience.rpi.edu/svn/public/ontologies/STOM/trunk/).

We continue to use the Inference Web explanation framework and its Interlingua for provenance, justification, and trust representation, PML - the Proof Markup Language, the Open Provenance Model, and associated tools. Given our maturing experience with these languages and tools, we have provided feedback and suggestions for improvements and additions to the IW effort and are participating in the World Wide Web Consortium’s (W3C) Provenance Working Group. We also continue to work with the University of Michigan on user presentation and integration of the tools we have developed into data product pipelines beyond those administered at HAO.

Key outreach and broader impacts continue with NASA Goddard Space Flight Center and the Australian CSIRO ICT Centre.

**Activities and Findings**

During the current period we have determined the following major findings:

- **Area 0) Foundation activity** - multi-domain knowledge bases and provenance granularity considerations.
- **Area 1) Use case refinement**, based on CoMP and discussions with HAO staff this summer.
- **Area 2) A provenance/metadata needs questionnaire**, aimed to get feedback from the solar physics community to help drive our work.
- **Area 3) Updated data ingest routines** to generate records for CoMP data products, encoded in RDF.
- **Area 4) Recording of provenance** for MLSO observer log entries.
- **Area 5) RDF record designs** for CoMP’s data products.
- **Area 6) A prototype implementation** for presenting these RDF records, based on supporting Semantic Web technologies (e.g., OpenLink’s Virtuoso triple store, SPARQL and XSLT).
• Area 7) Provenance triple store population, query and free text search

In the following sections, we elaborate on the abovementioned activities and findings.

**Area 0 - Multi-domain Knowledge Bases and Provenance Mapping**

Among the important findings from the previous year (figures and narrative not repeated here) was that answers to use cases queries come from knowledge bases that contain contributions from provenance, domain terminology, data product/processing terminology taking advantage of each context and multi-modal individuals in the classes for each of the ontologies. This is a fundamental aspect now of all our new activities and findings and will be a featured topic in the upcoming papers.

**Area 1 - Use Case Refinement based on CoMP**

This year, we continued to implement and refine some of the use cases reported last year as a result of the iterative development methodology and evaluations from the prototyping process. We only report on updates to the original use cases since last year first and then go on to report on the updates and refinement based on the new instrument CoMP.

• Use Case 2: Event Metadata
  • scenario: Provide details on observing conditions and observer comments surrounding the north-east limb event observed by PICS on March 25, 2009.
  • status: complete
• Use Case 5: Annotation Source Metadata
  • scenario: Who (person or program) added the comments to the science data file for the best vignetted, rectangular polarization brightness image (http://download.hao.ucar.edu/2005/01/26/20050126.184909.mk4.rpb.vig.jpg and http://download.hao.ucar.edu/2005/01/26/20050126.184909.mk4.rpb.vig.fts.gz) from January 26, 2005 18:49:09UT taken by the ACOS Mark IV polarimeter?
  • status: incomplete
• Use Case 6B: Process Source Metadata
  • scenario: What images comprised the flat field calibration image used on January 26, 2007 for all ACOS Mark IV polarimeter images?
  • status: complete
• Use Case 6C: Processing Stages Metadata
  • scenario: What processing steps were completed to obtain the ACOS CHIP quicklook images between September 18, 2008 at 18:34:06UT?
  • status: complete
• Use Case 7: Structured Query
  • scenario: Retrieve a JPG quicklook image under cloud-free (clear) conditions on Jan 26, 2005 for the MLSO CHIP intensity measurement and show the supporting evidence
Based in part on the above use cases, developed for the HAO Chromospheric Helium I Imaging Photometer - CHIP, five use cases were developed for use with CoMP earlier this year (use cases 1-5). Likewise, two additional use cases (use cases 6-7) were developed based on discussions with HAO Staff (in particular, Steve Tomczyk). Below we include an assessment of use cases 1-5, following discussions with Leonard Sitongia and Steve Tomczyk (the primary developers of CoMP's processing pipeline). In general, two concerns were raised on recording fine-grained provenance in CoMP:

(i) Need by HAO's consumers:
In general, CoMP's data processing workflow wasn't meant to be reviewed by anyone outside NCAR. Both Steve and Leonard mentioned that tech reports have been published on what CoMP does (e.g., http://www.cosmo.ucar.edu/publications/CoMP_preprint.pdf), which highlight the important processing steps carried out by CoMP. However, they felt that details beyond this (which highlight low-level system processing) may not be as useful for scientists.

(ii) Long-term support:
Steve Tomczyk mentioned that CoMP's code is still a work in progress, since the instrument recently went online. As such, new processing types may be added and process ordering may change, which could require repeated maintenance of fine-grained logging routines down the road.

Use Case 1 (adapted from CHIP):

Title: Explain generation of facts about a particular data product

Introduction: In CoMP, processed images are assigned quality scores (e.g., the Good-Bad-Ugly (GBU) metric), to aid scientists in obtaining high-quality products. Often, desired quality criteria will be relative to a scientist’s needs. Therefore, the ability to explain processing steps for quality metrics can be helpful. Such explanations may include facts like: processing inputs/outputs, process description, and processing timestamps.

Scenario: Alice is a solar physicist working outside HAO. She has no familiarity with the CoMP architecture, but expert-level knowledge of the corresponding domain. She wants to familiarize herself with how the GBU processing metric works, prior to downloading images graded as high-quality by HAO. This initial familiarization will require a list of processes applied to calculate the metric (with corresponding descriptions).

Assessment after HAO discussions:
Addressing this use case can be done by listing select facts about the GBU metric, as opposed to providing a detailed PML trace of the process. Facts could include:
- A description of what the GBU metric is
- A description of the calculations used in GBU
- A list of corresponding parameters used by these calculations, with accompanying values and descriptions

**Use Case 2 (adapted from CHIP):**

**Title:** Show/describe the derivation provenance for a published data product (i.e., a coronal intensity GIF).

**Introduction:** Here, the goal is to show a listing of processing algorithms applied to derive new images from the original ones obtained at MLSO. Here, data dependencies/outputs of processing algorithms will be omitted from presentation until specifically asked for.

**Scenario:** Alice is a solar physicist working outside HAO. She has no familiarity with the CoMP architecture, but expert-level knowledge of the corresponding domain. Prior to using any particular CoMP solar visualization in her work, she wants to familiarize herself with (or possibly verify) the steps taken by HAO to generate it. This review by Alice will require a list of processes applied to generate the solar visualization. Here, data dependencies/outputs of processing algorithms can be omitted from presentation until specifically asked for.

**Assessment after HAO discussions:**

Here, exposing a list of processing activities applied to derive a data product (as opposed to fine-grained provenance) could be useful as a reference to scientists. Facts about these activities could include:
- A description of the activity.
- Files used/generated by the activity.

It may be possible that scientists will want to know how certain processing was implemented in the pipeline code. To help with this, references could be provided from listed processing activities to source code used to run them. This could be done with the following content:
- A link to HAO’s SVN repository, along with an SVN path for the specific code file.
- An SVN revision number, indicating the code version used during pipeline execution.
- One or more contact persons, who developed the code and could aid consumers in a code review.

**Use Case 3 (adapted from CHIP):**

**Title:** Detailed explanation a single processing step in the derivation provenance for a published data product
**Introduction:** Here, the goal is to show as much detail as possible (without overwhelming an end user) corresponding to a single processing step in the derivation provenance from use case 4.

**Scenario:** For a particular solar visualization, Alice wants to know more about the corresponding flat-field calibration process. Specifically, she wants to know: (i) the algorithm used by this process, and (ii) the data products used by the algorithm.

**Assessment after HAO discussions:**

Currently, publications exist which highlight types of processing in the CoMP pipeline ([http://www.cosmo.ucar.edu/publications/CoMP_preprint.pdf](http://www.cosmo.ucar.edu/publications/CoMP_preprint.pdf)). It remains unclear whether reviewing lower-level computational details would be as useful for HAO’s data consumers.

However, details on algorithms applied to generate a data product could be tied to specific processing activities (see Use Case 2).

**Use Case 4 (adapted from CHIP):**

**Title:** Search for Observer Log entries by text search

**Scenario:** Show all observer log problem comments from 2001 that include the text 'tophat'.

**Note:** Applicable to both CoMP and CHIP, since both are tied to a common observer log from MLSO.

**Assessment after HAO discussions:**

In general, no concerns were raised on this use case.

**Use Case 5 (adapted from CHIP):**

**Title:** Search for observer comments by timestamp

**Scenario:** Show all observer comments from the local morning of January 26, 2005 at MLSO. This use case could be extended to address other observer log comment properties (e.g., authorship, comment category), possibly with multivariable search.

**Note:** Applicable to both CoMP and CHIP, since both are tied to a common observer log from MLSO.

**Assessment after HAO discussions:**

In general, no concerns were raised on this use case.
Use Case 6 (developed with Steve Tomczyk and Laurel Rachmeler @ HAO):

**Title:** Enabling scientists to obtain data surrounding a major event.

**Scenario:** Alice is a Solar Physicist working outside HAO, interested in CoMP’s readings of the solar corona. She is interested in obtaining a set of Stokes Intensity Images generated by CoMP, taken within 3 hours of a Coronal Mass Ejection occurring at a specified date and time.

Use Case 7 (developed with Steve Tomczyk and Laurel Rachmeler @ HAO):

**Title:** Enabling scientists to check for data availability, based on specified parameters.

**Scenario:** Alice is a Solar Physicist working outside HAO, interested in CoMP’s readings of the solar corona. She is interested in obtaining a listing of days during June, 2011 in which image data was gathered with a line center of 1083.

Based on these concerns, use cases 1-5 were modified to focus on the retrieval of a high-level provenance trace, which in turn can be annotated with processing details relevant to scientists. As an example from CoMP, when the main processing script (Demod.pro) is executed, processing activities such as Flat and Dark Correction (reported on in last year’s annual report) are applied to corresponding data. Rather than log a detailed PML trace of Demod.pro, we could instead represent Demod.pro as a "black-box", which can be annotated with processing activities HAO staff see as important.

**AREA 2 - Provenance/metadata questionnaire**

To get a better idea of what kinds of provenance/metadata Solar Physicists need/use, we developed the attached questionnaire. This version was reviewed and updated with feedback from Leonard Sitongia and Joan Burkepile. We made this survey available online at https://spreadsheets.google.com/spreadsheet/viewform?formkey=dHJYeHphNHFZaThEWlNsYzBqSFNuUnc6MQ, and announced through a few email listings (HAO's Staff Notes Daily and SolarNews lists, as well as the BESSIG's (Boulder Earth and Space Science Informatics Group) mailing list). Responses are still coming in and will be reported on in future reports/papers. The full survey is included at the end of this report.

We are also considering re-announcing the survey later on, if we still need this information.

**AREA 3 - CoMP**

As a result of the activities and findings in Area 2 and the discussions with HAO staff, we decided to move toward lightweight provenance logging/data ingest strategies, due to the concerns Leonard and Steve raised earlier (see above).
The logging strategies we've considered:

1) WDoIt: We initially considered this (in 2010) for managing provenance logging and modeling the CHIP pipeline. However, we were unable to get the program to work reliably.

2) IDL logging API: This was developed to allow provenance statements to be logged from CHIP/CoMP IDL code. The first version was developed as a C-based library, designed to be called from IDL. However, HAO staff pointed out this version of the API may cause conflicts with existing pipeline routines. The second (current) version is completely based on IDL routines.

Initially, our logging objective was to record as much information as possible, with the possibility that someone may want to see it later. However, this strategy changed based on feedback from HAO staff.

In its current version, our logging API is designed to support a mixture of: (i) high-level provenance recording, and (ii) annotation of high-level processes with kinds of processing activities occurring. In taking this approach, any maintenance of logging routines by HAO staff would be greatly simplified (versus maintaining logging routines for generating a low-level PML proof).

**Area 4 - Recording of provenance for MLSO observer log entries**

![Figure 1 Original observer comments presentation.](image-url)
This was a relatively short exercise to generate PML from observer logs using a simple java program (using the Jena API) that generates PML for observer comments from an MLSO observer log. The observer logs are scrapped from 2005 through part of 2010 (i.e. what is available at the time online from http://download.hao.ucar.edu/d5/mlso/log/observer/{ccyy}/mlso.{ccyy}.d{doy}.olog) and the most recent generated instance contains approximately 1926 PML documents.

The 'registry' root for the PML and associated RDF is located at http://escience.rpi.edu/2010/mlso/.

The registry has the following subdirectories

PER - RDF records of individuals referenced in the PML. Uses FOAF, MLSO, and PML predicates.
SEN - RDF records for sensors referenced in the PML. Uses MLSO, and PML predicates.
schemas - contains mlso.owl (not the same as the mlso.owl ontology from VSTO) and mlsoLog.owl ontologies reated for this exercise
PML - contains the generated PML records.

The purpose of this task was to structure and encode observer comments in RDF so they may be queried and presented in accordance with observer comment presentation section from the SPDIS mockup (image attached).

The exercise was a success, RDF/PML encoding successfully created and seemed simple enough to query and use, but progress on the presentation prototype stalled so we have not made much use if it yet.

Next steps are to review this PML and assess if it can be immediately integrated with the presentation prototype being developed, assessing if changes may have to be made. We also need to integrate with the mlso.owl and mlsoLog.owl schemas originally developed and adapt them to the new STOM work. Once complete we will be able to re-run the generation scripts and put the generated RDF into a triplestore; then add the observer comment section to the presentation prototype.

**AREA 5 - RDF record designs for CoMP's data products**

The current RDF design for CoMP's intensity filtergrams (so far, this is the only data product CoMP is producing) is shown in Fig. 2. This design covers over the following things:

1) Observation details - which link the data product to a CoMP observation activity. This leverages the STOM, O&M and VSTO ontologies.

2) Data quality - this helps describe quality metrics applied by HAO to data products. Here, the DQ ontology is used.

3) Provenance - this provides an overview of steps taken to derive the data product. Currently, OPM is used in the design, but could be replaced with PML (see below).
4) Processing Activities - a form of provenance annotation, used to describe activities within a process. Here, a processing activity is said to be "part of" a general process.

Figure 2. RDF design worked for CoMP's Intensity Filtergrams, which incorporates concepts from VSTO, STOM and DQ.

The next step in this area is to complete the provenance model representation evaluation and determine if a shift to PML would be preferable. From an expressivity perspective, either OPM or PML would work for this design. This evaluation will be reported in upcoming papers.

**AREA 6 - A prototype implementation for presenting these RDF records, based off Virtuoso, SPARQL and XSLT**

This area addresses the implementation strategy for storing and presenting RDF to HAO's data consumers. We will be storing RDF in a Virtuoso instance, hosted on an HAO machine. This triplestore will be populated by a Java application (based on Jena), designed to write the RDF records themselves (see Area 5). We expect this implementation to go online this September.

In turn, these records will be retrieved through SPARQL querying, and transformed via XSLT to a presentable form. Fig. 3-5 shows 3 screenshots of our current prototype, each highlighting different kinds of information the UI can present.
Image Details

Figure 3. Prototype presentation view for CoMP with main image details and initial metadata and further presentation options.

Processing (Expand)

Quality (Expand)

Header Entries (Expand)

Figure 4. Prototype presentation view with expanded detail on processing history.

Quality (Collapse)

Figure 5. Prototype presentation view with detail on quality information.
**AREA 7 - Provenance triple store population, query and free text search**

Based on one of the original use case - [http://tw.rpi.edu/web/node/910](http://tw.rpi.edu/web/node/910), which required exploration of free text search capabilities in triple stores. We also migrated PML data at [http://escience.rpi.edu/2010/mlso/PML/](http://escience.rpi.edu/2010/mlso/PML/) to triple stores. To this effect, we loaded PML data into TDB, Allegrograph and Virtuoso and tested free text search capabilities and performance in Virtuoso and Allegrograph. The results of this and the PML triple store information are documented at [http://tw.rpi.edu/web/project/SeSF/WorkingGroups/Triple_Stores/Triple_Stores_Evaluation](http://tw.rpi.edu/web/project/SeSF/WorkingGroups/Triple_Stores/Triple_Stores_Evaluation).

Our next steps in this area are to query provenance from triple store and this work is in progress. We have been using the Linked Data API (LDA) specifications ([http://code.google.com/p/linked-data-api/](http://code.google.com/p/linked-data-api/)) to implement an API to achieve this. There are a couple of implementations currently available namely Elda ([http://elda.googlecode.com/hg/deliver-elda/src/main/docs/index.html](http://elda.googlecode.com/hg/deliver-elda/src/main/docs/index.html)) and Puelia ([http://code.google.com/p/puelia-php/](http://code.google.com/p/puelia-php/)). We have begun working with Elda to implement an API but found it was not a complete implementation of the LDA specs. We are now exploring Puelia and will work on an implementation based on either Puelia or Elda depending on the functional outcome of the test.

**Next year’s plans**

Ongoing work from prior years includes:

- encoding of rules in the data workflow, encode those rules in OWL 2 RL (recently released) and configure Jena rule engine to support inference
- evolution of VSTO ontology (i.e. SESF/STOM/VSTO) to support provenance, service and rule concepts and attributes/annotations
- next steps for several of the areas list above

The expected final year work plan is:

- present and publish updated SESF = STOM+VSTO+SPCDIS ontologies, including provenance, services and rules
- complete implementation for CoMP instrument working at HAO
- present and publish current progress
- (objective) capability demonstration at regional or national science meetings
- work with SESF for full integration of SPCDIS within SESF data framework
- final report and final publications

**Participants, Management and Coordination**

PI Fox provides overall project management and coordination, expertise in use case development (and is PI for the VSTO), architecture development as well as data science expertise. The RPI team consists of Peter Fox (PI), Stephan Zednik (software engineer) and Patrick West (software engineer), Cynthia Chang (programmer), James Michaelis (graduate student), Mandeep Singh (masters student), Arun Prakesh Lakshmi (masters
student) and Rajashree Deka (masters student). Not all RPI personnel are fully funded on
this project. The project wiki documents progress on the effort and serves as a repository
for materials, such as reports, presentations, and use cases. Deborah McGuinness (co-I)
and James Benedict (research associate) from McGuinness Associates led the knowledge
representation and inference effort. Thomas Zurbuchen (co-I) from the University of
Michigan represents community interests and creation of higher-level data products
anticipated for the NCAR CoMP (Coronal Magnetic Polarimeter). We continued to
partners, from the University of Texas at El Paso including Paulo Pinheiro da Silva
(UTEP CyberShare Center, Project CO-PI), Nicholas Del Rio, Aida Gandara, Leonardo
Salayandia, and Jitin Arora – all UTEP Computer Science, PhD students.

Conference

Meetings with SPCDIS representation, presentation.

The project team has had several in-person meetings (as well as several teleconference
calls and numerous email correspondence):

• Project meetings (at RPI, approximately every two weeks - wiki
  http://tw.rpi.edu/wiki/Provenance through Dec. 2010 and then
  http://tw.rpi.edu/web/ResearchAreas/KnowledgeProvenance afterwards)
• AGU Fall Meeting 2010 (team meeting and provenance talks by Zednik, West)
• RPI collaboration visits to NCAR/HAO, August 31st, 2010 and January 15th,
  2011 (Michaelis, West, Zednik)
• RPI summer intern at HAO (James Michaelis)
• RPI collaboration visit to UTEP, Feb 2011 (Zednik and West), April 2011 (Fox),
  UTEP collaboration visit to RPI, June 2011 (Pinheiro da Silva)
• CSIRO collaboration visit, October 2010 (Zednik and West)
• CSIRO collaboration visit, June 2011 (Fox)
• European Geosciences Union (EGU) General Assembly 2011, April 2011
  (Zednik)

Planned events for the coming year:

• Project meetings (RPI)
• UTEP collaboration visit, planned for November 2011
• Inference Web - RPI/UTEP workshop October 2011
• CSIRO collaboration visit to RPI, September 2011 (David Lemon)
• CSIRO collaboration visit, November 2011
• Periodic NASA/GSFC collaboration visits, September 2011 through February
  2012
• Semantic Web for Provenance Management (SWPM) at ISWC 2011
• AGU Fall Meeting 2011 (and team meeting), December 2011
• EGU General Assembly 2012, April 2012
Publications and Products


Contributions

Education

Four graduate students at RPI and four at UTEP are participating in this project.

Outreach/ Impact

We continue to have significant outreach and broader impacts on other data ingest systems and workflow pipelines efforts this year with the NASA Earth Science Technology Office (ESTO) project entitled Multi-sensor Data Synergy Advisor (MDSA with PI Greg Leptoukh), and two NASA ACCESS projects; the Data Quality Screening Service (DQSS with PI Chris Lynness) and an Environment for Assessing the Status of Aerosol Measurements (AeroStat with PI Greg Leptoukh). See the respective project wikis for details - [http://tw.rpi.edu/web/project/MDSA](http://tw.rpi.edu/web/project/MDSA), [http://tw.rpi.edu/web/project/DQSS](http://tw.rpi.edu/web/project/DQSS) and [http://tw.rpi.edu/web/project/AeroStat](http://tw.rpi.edu/web/project/AeroStat).

We are active in a formal research collaboration agreement with the Australian CSIRO ICT Centre (in Hobart, Tasmania) in the area of Provenance Management for Hydrological Sensor Network as part of their water informatics efforts ([http://tw.rpi.edu/web/project/SWaMP](http://tw.rpi.edu/web/project/SWaMP)). In 2010-2011 CSIRO has expanded its provenance activity into 4 new project and RPI is involved in all of them.

The continued recognition by these two major institutions of our progress and impact, in completely different disciplines confirms our broader impacts arising from this project.

Lastly, the planned modification to the community ontologies developed for the SESF in the area of domain aware provenance tools will mean a sustained place for them ongoing development, and dissemination beyond this project.

Distributed Registries

As part of leveraging the existing tools and software and data sources, we have a fully distributed capability from the start. Specifically, we have the following sites and capabilities:
Special Requirements
All students have completed the Responsible Conduct of Research (RCR) training.

SUMMARY
We continue to meet our objectives and are proceeding in the final year to complete the project and merge the results into the SESF project.

Appendix A – Activity Area 2 – Questionnaire

A Research Survey on Provenance Usage by the Solar Physics Community

This research survey is being distributed by the Tetherless World Constellation (TWC - http://tw.rpi.edu/) at Rensselaer Polytechnic Institute (RPI - http://www.rpi.edu) in its ongoing research into the collection of knowledge provenance metadata (SPCDIS - http://tw.rpi.edu/web/project/SPCDIS). This research is being conducted by James Michaelis, Ph.D. student at RPI, under the guidance of Professor Peter Fox. For more information, please contact James at michaj6@rpi.edu. TWC would like to thank Leonard Sitongia and Joan Burkepile, researchers at the NCAR High Altitude Observatory, for their assistance with this survey and project.

Introduction:

The NCAR High Altitude Observatory (HAO) (http://www.hao.ucar.edu/) currently publishes data products (e.g., visualizations of coronal activity) online for consumption by interested parties. For this survey, we seek to identify patterns in usage of provenance metadata by members of the Solar Physics research community. In doing so, we hope to obtain additional insight into how provenance metadata can, for instance:

- Facilitate search for data products, such as solar visualizations.
- Enable data product retrieval based on defined quality metrics.

Provenance Metadata:

Definition: A record that describes entities and processes involved in producing and delivering or otherwise influencing a (physical or digital) resource. Provenance provides
a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance.

**Example:** For a particular cake, a corresponding provenance record may look like the one below (Figure 1). Here, the cake can be viewed as a type of *artifact*, created through a sequence of *processes* (Mix, Bake), each relying upon on alternate artifacts. Often, the ability to distinguish the role (or purpose) of each artifact in a process can be useful. In this case, to distinguish cake ingredients (e.g., sugar) from tools and appliances (e.g., mixing bowl).

![Diagram](image)

*Figure 1. Adapted from figure 3, page 5 of Moreau, L. et al. (2010) The Open Provenance Model core specification (v1.1). Future Generation Computer Systems. (In Press)*
Personal Info

Name: ______________________________________________________
Email: ______________________________________________________
Organization: _____________________________________________
Role(s) in Organization: _____________________________________
Area(s) of Domain Expertise: _________________________________

Experience with:

Workflow management programs:
[ ] Kepler (http://kepler-project.org/)
[ ] Taverna (http://www.taverna.org.uk/)
[ ] Other (please list _____________________________ )

Data Analysis Suites:
[ ] IDL (http://www.ittvis.com/)
[ ] Matlab (http://www.mathworks.com/products/matlab/)
[ ] Other (please list _____________________________ )

Metadata Languages:
[ ] ISO 19115
[ ] Dublin Core
[ ] OGC Observation and Measurements (O&M)
[ ] CF Conventions
[ ] Other (please list _________________________ )

Metadata Formats:
[ ] XML
[ ] Microformats (RDFa, Microdata, etc.)
[ ] RDF (Resource Description Framework, http://www.w3.org/TR/rdf-syntax/)
[ ] OWL (Web Ontology Language, http://www.w3.org/TR/owl2-overview/)
[ ] FITS Header Record
[ ] CDM Attributes
[ ] Other (please list _________________ )
Survey Introduction

Please provide information based on any interaction(s) with digital systems as part of your work. Simple use cases, driving your interaction with these systems, would be helpful.

Here, use cases are defined as descriptions of steps or actions between a user and a software system, leading the user to accomplish a particular goal. In the context of NCAR, an example use case could be as follows:
At HAO, raw data gathered from the Mauna Loa Solar Observatory (MLSO) is processed into data products usable by scientists (e.g., visualizations of coronal activity), and assigned quality scores based on various metrics (e.g., levels of detected image noise). One metric, known as GBU, assigns an image grade of ‘Good’, ‘Bad’, or ‘Ugly’, based on automated algorithms.
Alice is a solar physicist working outside HAO, interested in data products derived from MLSO’s Coronal Multi-Channel Polarimeter (CoMP). She has no familiarity with the HAO’s processing techniques, but expert-level knowledge of the corresponding domain. She wants to familiarize herself with how HAO’s GBU processing metric works, prior to downloading images from the HAO website graded as high-quality. To do this, she obtains from the website a list of algorithms applied to calculate the metric for CoMP-derived data (with corresponding descriptions).

Questions

Are you familiar with any provenance languages? If so, please list.

What provenance metadata are you most likely to use?

What provenance metadata most helps you determine if you trust a product? (here, trust is defined as the willingness of one party (the consumer) to rely on a product generated by another party (e.g., a human or digital system)
How do you use quality assessment metrics (when available) in your processing of published data products?

What provenance questions would you be most likely to ask when assessing a data product? (examples - How was this image’s quality assessment value determined? What original observations went into the creation of this daily map? What processing was done to generate this product? Who performed the processing on this product? What calibration technique was applied to this product? What inputs were used to generate this data product’s calibration image? etc.)

What provenance metadata are you most likely to use as search criteria?

Would you use provenance metadata if it was provided out-of-band? (i.e. not included in the data product itself, perhaps available via a URL. The provenance URL may be included in the data product metadata)
Would you be likely to use provenance metadata if it was not packaged within the data product? (i.e. out-of-band XML or RDF metadata for a FITS data product)