System Transparency, or How I Learned to Worry about Meaning and Love Provenance!

Stephan Zednik, Peter Fox, and Deborah L. McGuinness

Rensselaer Polytechnic Institute, Troy NY 12180, USA

Abstract. Web-based science analysis and processing tools allow users to access, analyze, and generate visualizations of data without requiring the user be an expert in data processing. These tools simplify science analysis for all science users by reducing the data processing overhead for the user. The benefits of these tools come with a cost, the increased need for transparency in data processing. By providing a clear explanation of the science concepts and processing performed by the science analysis tool we can increase user trust, understanding, and accountability and reduce misinterpretation or generation of inconsistent results.

We will demonstrate knowledge provenance (processing lineage and related domain information) presentation capabilities applied to an existing web-based Earth science data analysis tool (e.g. Giovanni from NASA/GSFC). Our conclusion is that user accessible visual presentations of knowledge provenance are key to building meaningful user understanding of analysis and processing decisions and should be a key component of data analysis tools.

1 Introduction

Science communities are putting increasing emphasis toward sharing data and developing publicly accessible tools to support streamlined analysis and visualization of this data. These tools are of great benefit to the community, as the burden of dealing with downloading data, accessing specialized data formats, running analysis processes, and complicated plotting tools is lifted from the user, allowing them to get directly to the core of their research. These powerful user tools come with a hidden cost; while the barrier to entry is lowered since the user does not have to manually address system-specific behaviors of the analysis operations[1], the user may also be unaware of a multitude of system, science, and data lineage details that can negatively impact the scientific or statistical applicability of the results.

We aim to develop a multi-function provenance system geared toward enhancing user understanding of data products and information derived from science analysis tools. User-oriented visual presentations of science knowledge and processing provenance represent the key functional requirement to achieving our goal. To achieve a reasonable level of understanding regarding the fitness for purpose of science data, a user should be aware not just of what processes were run

[1] Data access, format translations, data calibrations and screenings, etc.

D.L. McGuinness, J.R. Michaelis, and L. Moreau (Eds.): IPAW 2010, LNCS 6378, pp. 165–173, 2010.

to produce the data product, but the science intent of the processing and science concepts associated with processing and data throughout the processing trace. We call this integration of processing history and science concepts, knowledge provenance, and we believe it is integral in developing transparent, open science applications. Beyond just attempting to capture this knowledge provenance, we must present it to the user in a manner designed for human consumption, yet thorough - with hooks that allow the user to dig into the web of knowledge and follow concepts to their definition and, ideally, provide understanding. It is our assertion that exposing rich knowledge provenance to the science tool user in a manner that is cognitively pleasing to use, easy to navigate and informative of meaning, will significantly enhance user understanding of science data and the processes used to develop it.

We illustrate our work toward knowledge provenance representation and presentation for operation of a test environment of the NASA Giovanni [1,2] interactive online Earth science data visualization and analysis tool. Giovanni allows Earth scientists, interdisciplinary and other applications researchers to perform multi-sensor and model data analysis online, e.g., explore connections between atmospheric processes and sea or land surface properties. Giovanni is a publicly available production tool that is actively used by modelers, researchers, application users, policy makers, teachers, and students.

2 Use Cases

Our initial use case revolves around capturing and presenting Giovanni processing provenance with the specific goal of exposing this provenance for the user's visual consumption. Further use cases expand upon our presentation of the knowledge provenance to include highlighting potential differences in the knowledge provenance lineage of two compared products and advising the user on potential negative applicability factors in a data comparison by analyzing the knowledge provenance of the compared artifacts.

Provenance Visual Lineage/Proof Use Case: Provide a visual representation of processing and knowledge provenance for a time-averaged latlon map comparison of Aerosol Optical Depth from MODIS Terra and Aerosol Optical Depth from MODIS Aqua over the calendar period of 2008-01-01 to 2008-01-31.

Our basic provenance use case; capture and visually present the provenance to the end user. This scenario does not perform analysis of the provenance - just capture and presentation. Knowledge provenance is presented as a causality graph based on the provenance data lineage integrated with domain metadata.

Provenance-aware Advisor Use Case: Use Giovanni to compute a difference map of MODIS Daily Aerosols from Aqua and Terra Platforms, using knowledge provenance to

1. *understand the differences between the compared products*
2. *explain anomalies that may be present in the generated difference map.*

A more complex provenance use case; our system now provides applicability information, based on an analysis of the knowledge provenance, that a non-expert user would not necessarily glean from a raw visual presentation of the knowledge provenance. The basic flow for this use case scenario is:

System uses descriptive logic to determine when target domain concepts in the knowledge provenance are different in a manner that may affect a comparison

1. *System determines that the two datasets correspond to two different MODIS*[2] *sensors on two different satellites (Aqua (EOS PM-1) and Terra (EOS AM-1)).*
2. *System determines that the two satellites have different Nominal Equatorial Crossing Times (NEQCT) (13:30 for Aqua and 10:30 for Terra)*
3. *System determines that the two satellites have different daytime nodes*[3] *(Ascending vs Descending)*
4. *System uses these differences, together with the dataset DataDay definition*[4] *to infer that there is a difference in the local observation times included in grid cells in each product, with differences being greatest over the Central Pacific (see Figure 1.)*

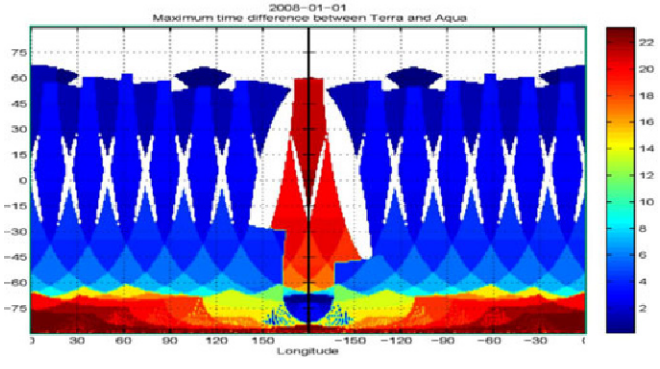


Fig. 1. Map of the time difference discrepancy between MODIS Aqua and MODIS Terra spatial coverage over the Pacific Ocean

We believe this is a critical provenance use case because it highlights how integrated provenance and domain information can be used by an intelligent system to discover highly-relevant information that would not be easily determinable if provenance and science domain metadata is disjoint.

3 System Requirements

The present production Giovanni service does not capture and retain provenance information, therefore, we developed a testbed Giovanni service based on the following architecture requirements

[2] Moderate-Resolution Imaging Spectroradiometer.
[3] Segment of the orbit transiting the Earth in daytime.
[4] A specification of how data are aggregated into daily data products, i.e., which pixels or scenes are included in a given day.

- develop test Giovanni environment to capture processing lineage
- encode processing lineage in a provenance interlingua
- encode domain metadata related to artifacts and processes referenced in the processing lineage in one or more domain interlinguas
- domain and provenance interlinguas must be integrable, that is to say the domain and provenance metadata should not be disjoint
- integrated knowledge base should be supported by standard query and rule systems
- visualization service should generate a user-orientated presentation of the knowledge provenance from the integrated metadata

To satisfy the provenance capture requirement our testbed Giovanni service produces a log of processes and inputs/outputs from processing. This log is translated to conform to an OWL[5] data lineage model. The artifact and processes identifiers in the log allow us to link the data lineage RDF graph with externally defined domain metadata encoded in OWL. By integrating the data lineage with the externally defined domain metadata we construct a knowledge base that supports mixed provenance / domain queries and reasoning (which can be used to infer domain or provenance information about entities in the knowledge provenance). Domain specific conditions related to provenance, such as highlighted in the provenance-aware advisor use case, can be checked by query or ruleset - with advisory or warning entities being declared in the knowledge provenance when the specific conditions are found. These rules and queries are engineered by domain experts but the advisories/warnings issued in the knowledge provenance contain descriptive metadata and become a part of the total knowledge provenance. Tools that support just the standard provenance OWL vocabulary can be used to generate a visualization of the data lineage; and tools that additionally support the integrated domain metadata can generate a visualization of the entire knowledge provenance.

4 Knowledge Provenance

We use the Proof Markup Language [3, 4] (PML) OWL ontology as a general-purpose provenance interlingua to encode a justification for the generation of Giovanni data visualizations. PML was chosen because it is a published provenance interlingua designed to encode justification metadata about general information or objects produced by an agent or decision mechanism. This generally scoped information-centric view of provenance lends itself well to our definition of knowledge provenance and we can make use of already-existing PML supported tools. Leveraging an OWL ontology as our provenance interlingua was also a supporting factor because OWL supports our requirement of using an interlingua system that supports the integration of multiple domain interlinguas and for which established query[6] and reasoning[7] mechanisms already exist.

[5] Web Ontology Language.
[6] SPARQL.
[7] SWRL, Jena Rules.

This mix of processing and science information in the provenance, which we call knowledge provenance, results in a very versatile knowledge base that supports a wide range of science-focused provenance reasoning. For example, in our provenance-aware use case, the applicability of a processing algorithm can be checked against the spatial and temporal resolutions of the service input dataset. Comparison integrity between two parameters or datasets can be checked based on a large number of factors, most encoded in the science metadata but reached by traversing the provenance lineage of the compared data products and their sources. These reasoning checks and queries would not be easy or straightforward if the science metadata and data lineage / processing provenance existed in disjoint knowledge bases.

5 Provenance Visualization

Probe-It! [5] is a provenance browser suited to graphically render provenance information encoded in PML. Probe-It! does not generate content, but renders an interactive visual representation of a provenance causality graph.

In Probe-It!, users can select nodes within the provenance trace to see a detailed view of the justification for the process/decision at that point. As of the time of this writing, Probe-It! only supports the presentation of information encoded in PML properties, so statements in the non-PML vocabulary are not visible to the user. Much of the science information in our knowledge provenance is not encoded in the PML interlingua, rather this information is stated in domain interlingua and related to the data lineage through entities common to both the PML graph and the domain graph. For the moment we are encoding the values of non-PML properties that capture important aspects of images, datasets, satellites, and processes into string description properties from the PML vocabulary. This gives us a basic mechanism to represent domain specific information (such as temporal and spatial resolutions of a dataset, or the orbital characteristics of the satellite an instrument operates on) from PML tools that do not support our domain interlinguas.

An early Probe-It! visualization of the provenance generated by our provenance visual lineage use case scenario is shown in Figure 2. The highlighted center node represents the process that extracts the requested data from the data source based on user selections of dataset, parameter, and spatial and temporal constraints. The output, or conclusion, of this process is a set of temporary data files listed in an XML fragment. This XML fragment can be viewed by the user using Probe-It! detail view of the node, or a human-friendly summary of the results of the process may also be shown. In this early representation of the knowledge provenance a fragment of the testbed Giovanni processing log detailing the output of the selected process are encoded in a PML string property as a representation for the conclusion of the process. This interface highlights the need for support of domain vocabularies in the visualization tools; both to support better domain presentation as well as to retain semantics of the domain interlingua and entities.

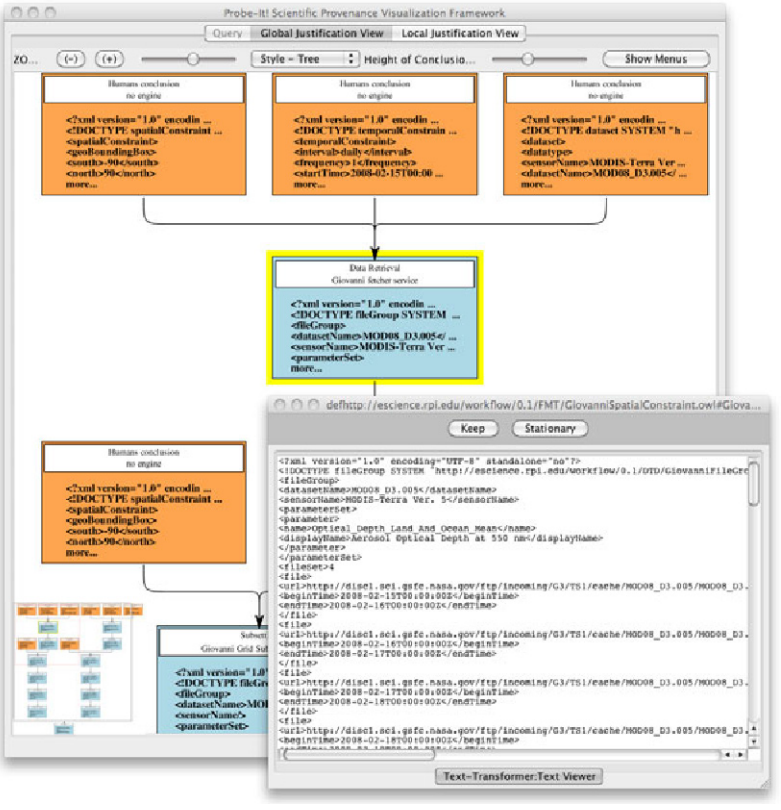


Fig. 2. Probe-It! visualizing the provenance of a data extraction process during a Giovanni data analysis comparison

We have been working with the Probe-It! team at UTEP[8] to determine the best way to present information from non-PML vocabularies (our domain interlinguas). Changes made to Probe-It! that support presentation of general vocabularies are folded back into the core Probe-It! browser.

The Probe-It! visualization is made accessible to the Giovanni user by way of a link that is available on the Giovanni results page. When the user clicks this link the Giovanni service invokes the Probe-It! web applet, passing to Probe-It! the entity URI of the final conclusion from the provenance trace, and Probe-It! automatically loads the provenance trace for the passed in information.

Internal use and informal evaluation of Probe-It! found that while processing provenance structure is clear and easy to follow, science information encoded as non-semantic text in the PML description properties can be hard to understand and difficult for users to act upon. User analysis of the provenance trace to

[8] University of Texas at El Paso.

determine the similarity of compared products or to discover potential applicability issues with processing actions was found to be especially difficult.

Subsequently, a visualization was defined by domain experts from the Giovanni team, whereby selected science information in the knowledge provenance is highlighted for the user in a concise table, geared towards the comparison and analysis scenarios for which a generic lineage presentation was found difficult to use. The table has a column for each data selection in a comparison analysis, to support a clear and simple presentation of differences in the knowledge provenance within the data selections. A set of simple semantic (Jena) rules was developed to search for semantically important differences in the knowledge provenance and if found, an advisory is generated in the model data displayed within the table with links to descriptive information regarding the posted advisory. These rules have been extended beyond simple semantic differences to include complex scenarios where multiple science aspects of the artifacts, along with certain processing actions, lead to anomalous results in the data visualizations. An example presentation of the domain differences found and advisory issued based on our provenance-aware advisor use case is shown in Figure 3.

The table representation of knowledge provenance has tested very well with our internal science group. Its ability to show, side-by-side, discrepancies between knowledge provenance of compared artifacts along with advisories and warnings about the comparison has proved to be an improved way to relate actionable information to the end user.

We plan to continue work on both (browse and table) visualizations of provenance. Both presentations have significant strengths, and both have areas where clarity or detail could be improved. At present, we do not know if these presentation scenarios will converge, but co-development should bring significant improvements to both.

6 Demonstration

We plan to demonstrate at IPAW 2010[9] execution of our provenance-capturing testbed Giovanni service and both visual presentations of the resulting provenance. We will show how Probe-It! is used to present a provenance trace of the Giovanni service execution and how Probe-it!'s local view can be used to access both processing and domain metadata about nodes within the provenance trace. We will also show how the Giovanni table view of knowledge provenance is used by domain experts to highlight differences in selected factors related to the provenance of compared data products and how the table view is used to inform users of comparison advisories and warnings in their data selection.

7 Discussion and Conclusions

To date, we have developed the ability to generate processing and science knowledge provenance for execution of a test environment of the NASA

[9] The Third International Provenance and Annotation Workshop, `http://tw.rpi.edu/portal/IPAW2010`

Your Selected Options:

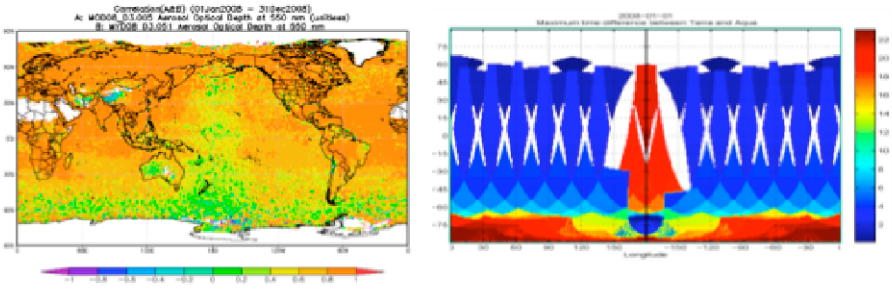
Spatial Area:	Longitude (-30, 150), Latitude (-10,60)
Parameters:	A: MYD08_D3.005 Aerosol Optical Depth at 550 nm B: MOD08_D3.005 Aerosol Optical Depth at 550 nm
Temporal Range:	Begin Date: Jan 01 2008 End Date: Jan 31 2008
Visualization Function:	Lat –Lon map Time-averaged

About your selected parameters:

	Parameter A	Parameter B	Difference alert
Parameter Name :	Aerosol Optical Depth at 550 nm	Aerosol Optical Depth at 550 nm	
Dataset:	MYD08_D3.005	MOD08_D3.005	← Diff
Data-Day definition	UTC (00:00-24:00Z)	UTC(00:00-24:00Z)	The same but....
Temporal resolution	Daily	Daily	
Spatial resolution	1x1 degree	1x1 degree	
Instrument:	MODIS	MODIS	
Satellite:	Aqua	Terra	← Diff
EQCT	13:30	10:30	← Diff
Day Time Node	Ascending	Descending	← Diff
Pre-Giovanni Processes :	*ATBD-MOD-30*	*ATBD-MOD-30*	
Giovanni Processes:	Spatial subset Time average	Spatial subset Time average	

Known Issues:

The difference of EQCT and Day Time Node, modulated by data-day definition, caused the included overpass time difference, which makes the artifact difference. See sample images:



MODIS Terra vs. MODIS Aqua AOD Correlation

Included Overpass time Difference

Continue process to display image

Return to selection page

Fig. 3. Knowledge provenance table, and advisories, for the visualization comparison from the use case scenario

Giovanni interactive online Earth science data visualization and analysis tool. This knowledge provenance captures both processing and science concepts involved in artifacts/information and processing within the system. Two user focused presentations of this provenance have been utilized.

- Probe-It! is used to browse the causality graph of Giovanni processing and contains a simple representation of science information for the artifacts and processes in the provenance trace.
- A Knowledge Provenance Table is used to show properties from the knowledge provenance side-by-side when Giovanni is used to generate an analysis comparison. This mode has been shown to be useful in showing the end user if the data selection comparison is potentially invalid.

The next stage of our work will be to increase and refine the expressiveness of our knowledge provenance, increase support and utility in Probe-It! for presentation of information from non-PML vocabularies, and further develop the knowledge provenance table presentation. The results of this work will eventually be incorporated into the production Giovanni analysis tool.

Acknowledgments

- This work was supported in part by the NASA ESTO AIST-08-071 project "Multi-Sensor Data Synergy Advisor" (PI: Gregory Leptoukh, NASA GSFC[10])
- The PML group at Inference Web
- The Probe-It group at UTEP/CyberShARE

References

1. Acker, J.G., Leptoukh, G.: Online Analysis Enhances Use of NASA Earth Science Data, 2007. Eos, Trans. AGU 88, 14–17 (2007)
2. Berrick, S.W., Leptoukh, G., Farley, J., Rui, H.: Giovanni: A Web Service Workflow-Based Data Visualization and Analysis System. IEEE Trans. Geoscience and Remote Sensing 46, 2788–2795 (2009)
3. da Silva Pinheiro, 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)
4. 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)
5. Del Rio, N., Pinheiro da Silva, P.: Probe-It! Visualization support for provenance. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Paragios, N., Tanveer, S.-M., Ju, T., Liu, Z., Coquillart, S., Cruz-Neira, C., Müller, T., Malzbender, T. (eds.) ISVC 2007, Part II. LNCS, vol. 4842, pp. 732–741. Springer, Heidelberg (2007)

[10] Goddard Space Flight Center.