Provenance Artifact Identification in the Atmospheric Composition Processing System (ACPS)

Curt Tilmes
NASA Goddard Space Flight Center
University of Maryland, Baltimore County

Yelena Yesha
University of Maryland, Baltimore County

Milton Halem
University of Maryland, Baltimore County

Abstract

The Atmospheric Composition Processing System (ACPS) evolved from the heritage processing systems currently processing ozone data at NASA, Goddard Space Flight Center. The ACPS includes complete provenance tracking of the various artifacts related to data processing. These include the data transformation algorithms and all data in the system, both inputs from external sources and data produced within the system. Other artifacts include the hardware and software of the processing framework, the source instruments and satellites, scientific literature and documentation, and people and organizations. The origin of any data or algorithms is recorded and the entire history of the processing chains are stored such that a researcher can understand the entire data flow. Provenance is captured in a form suitable for the system to provide basic scientific reproducibility of any data product it distributes even in cases where the physical data products themselves have been deleted due to space constraints. This paper will discuss the identification of provenance artifacts in the system and web services for communicating metadata about those provenance artifacts.

1 Background

Earth Science research has evolved to require huge amounts of data and the application of complicated strings of data processing to distill actionable information from the vast amount of remotely sensed data. Capturing the provenance information from the entire processing chain is an enormous problem, but is a requirement for reproducibility which is the hallmark of science. Recent reports in the popular press have disturbingly highlighted cases where data have been lost or misreported and where scientists can't adequately convey information about their research. In particular, global warming has become such a 'hot bed' issue that many have become interested in following the paths of experts to either confirm or cast doubt on their research.

2 Object Identity

There are a multitude of identifier schemes, each with various advantages and disadvantages. The bioinformatics community has standardized on the Life Science Identify (LSID) and LSID Resolution System (LSRS) for a simple and elegant solution to the problem. [2] $^{\text{my}}$Grid uses that scheme for establishing identities of workflow provenance and data products. [4] In particular, the Taverna IDSet aggregation identity scheme has similarities to the ACPS DataSet concept, though the ACPS ArchiveSets and granularity iterator enumerations simplify aggregation for the special case of geospatial/temporal earth science data. We follow the *Strong Identification* and via our DataSets, the *Strong Identification with IDSet* identification strategies as described by Chapman and Jagadish. [1]

The NASA Earth Science Data Systems Working Group (ESDSWG) is currently drafting a study on identifier schemes and the Federation of Earth Science Information Partners (ESIP) Preservation Cluster has an identifiers test bed. Each of those efforts of focused exclusively on data files and data sets. See URL, URN, URI, PURL, DOI, OID, ARK, Handle, UUID, etc. for some options under consideration. I believe the earth science community will ultimately converge on something similar to LSID. As long as the scheme can be linked with a URI (all of these can), it will be easy to establish equivalences to objects referenced by the ACPS identifiers.

3 ACPS

The ACPS [5] has a particular focus on atmospheric composition (what are the particular constituents of the atmosphere?) and is the primary processing system for the total column ozone measurement at NASA. The heritage system was largely monolithic providing a very controlled processing flow from data ingest of satellite data to the ultimate archive of specific operational data products. The ACPS, however, allows more open access with standard protocols to various modules within the system, including an extended data archive, metadata searching, production planning and processing. This enables researchers to download publicly released versions of the processing algorithms and reproduce their processing remotely, while interacting with the ACPS. The algorithms can be easily modified allowing better experimentation and rapid improvement. The modified algorithms can be integrated back into the production system for large scale bulk processing to evaluate improvements.

3.1 External data access

We have also extended the system to include external access to some publicly available data sets from other instruments, including non-NASA missions as appropriate. The data we incorporate in this manner is typically available through direct WWW URL (HTTP or FTP), and we are able to execute production rules in our system to determine the appropriate data file and map it to the specific URL from which it can be obtained. (Some data sets are available on tape or some other mechanism requiring a time delayed order either for staging retrieval from tape, or entering into a queue for "process on demand" production. Some may even require a manual intervention. These can be treated as special cases of the same action.) We can record the URL (or, more generally, URI) as an identifier similar to the way we record identifiers for our own data, but what we find is that these identifiers are often not persistent. As science progresses and newer versions of algorithms or calibration result in reprocessing of data which changes the data to a later version. The older, obsolete data is typically removed from the archive. Even though such data may no longer be useful for new investigations, it may still a part of the provenance record of prior published research, and as such, identifiers should remain valid, and should clearly distinguish such old data from the 'latest' version.

3.2 Granularity

A key concern in establishing the identity of an object is the granularity of that object. Consider data as a hierarchy, from all data from all places for all times down to a single measurement of some property for a single place for a single instant in time. Dealing with data at either of those extremes of granularity is awkward. Convention breaks down data in "granules" somewhere along this hierarchy such that neither the size of a single granule nor the quantity of total granules are overwhelming.

A **Granule** is an individually identifiable portion of data. The dataset's **Granularity** refers to the differentiation between granules of the dataset, possibly in time or space. For the ACPS, a Granularity is a concrete entity, and has a well-defined mechanism for iterating through the member granules of the dataset. We also define a **Granule Key** as the distinguishing portion of a dataset identifier that select a specific granule from the dataset. Granules are usually, but not always, stored in a single file. For the purposes of this paper, assume that each granule is a single file.

For example, the Ozone Monitoring Instrument (OMI) Level 2 data are organized by orbit. For a Level 2 dataset, there is one granule (one file) per orbit. The Granularity is "Orbital" and the "Granule Key" is the orbit number.

Other typical granularities might be time based, such as each 5 minutes, daily, or tile schemes with X and Y coordinates that require nested iteration through space and time to enumerate the granules. The ACPS uses an RFC 2445[1] compliant iterator capable of complex date/time recurrences (i.e. Each third Wednesday of the month).

We group granules produced in a consistent way (e.g. from a consistent calibration, and by a specific version of the production algorithm, even though particular granules differ in other specific details (e.g. production date/time or production host) into **ArchiveSet**s. Within an ArchiveSet, there can never be two files with the same Granule Key at the same time for a given **ESDT** ("Earth Science Data Type"). The ACPS timestamps ArchiveSets so the particular set of granules (differentiated only by their Granule Keys) that were part of a given ArchiveSet at that time can be referenced by a **DataSet** identifier. They can be enumerated by executing the iterators associated with their Granularity.

4 Provenance Artifact Identifiers in the ACPS

For the ACPS we identify each artifact that contributes or is associated with the production of data within the system. This begins with the basic data flow artifacts that are most commonly considered (inputs and outputs), but continues from there to more detailed configuration management details of the system and even the agents responsible for controlling various events.

Here are a few of the artifacts currently being tracked:

- Basic Data Flow: Input Granules/Files, Input Parameters, Algorithm Name and Version, Output Granules/Files

- Detailed provenance of the "Data Transformation Event": Execution Environment (HW, SW, OS), Controlling Agent (Organization, Person)

- Detailed provenance of algorithm executable: Algorithm Source Code and Version, Algorithm Executable Version, Specific Build Artifacts (Compiler, Libraries, etc.)

- Detailed provenance of "Executable Build Event": Build Environment (HW, SW, OS), Controlling Agent (Organization, Person)

- Other Artifacts: Granularity (Orbital, Daily, Geographic Tile, etc.), Earth Science Data Types (ESDT), Algorithm, APP (integrated algorithm), Aggregated Dataset Identifiers

In the ACPS, we have both *generic identifiers* which are, in effect, metadata searches for objects and *specific identifiers* that resolve to an instance of the object, and often a physical ("bunch of bits") object. The specific object resolved from the generic identifier can change over time with calibration/algorithm changes or reprocessing with better ancillary data. Which identifier to use depends on the use of the identifier.

4.1 PURLs for ACPS artifacts

The ACPS has been experimenting with **P**ersistent **U**niform **R**esource **L**ocators (PURLs)[2] for assigning a specific identifier to each artifact. Each PURL is a **U**niform **R**esource **I**dentifier (URI), and also an actionable **U**niform **R**esource **L**ocator which can be resolved directly.

The general form of our identifiers includes an artifact type and an artifact identifier within that type. Depending on the type of artifact, some identifers further distinguish broad classes of artifacts into more specific classes hierarchically. This is a natural, obvious application of the URL syntax.

```
http://purl.org/NET/ACPS/<ArtifactType>
         /<ArtifactIdentifier>
```

`http://purl.org/NET/ACPS/Granularity/Orbital`
`http://purl.org/NET/ACPS/Granularity/Daily`

Granularity metadata describe the Granule Key, and the Iterators responsible for enumerating instances of that Granule Key.

`http://purl.org/NET/ACPS/APP/OMTO3/1.2.5`

Algorithm Plugin Package (APP) is a fully integrated algorithm, that includes all the information needed to execute the algorithm in the context of the ACPS. This includes default values for runtime parameters and static data (input files that remain constant across a dataset, regardless of Granule Key) and production rules for determining dynamic (i.e. change for each Granule Key) runtime parameters and input files.

`http://purl.org/NET/ACPS/BuildEvent/125526`
`http://purl.org/NET/ACPS/DataEvent/52782`

Various 'events' are assigned first class identifiers. They link other artifacts to the event that used or created them.

`.../Granule/17/OMTO3/28794`
`.../Granule/17/OMTO3d/2009-12-01`

These are 'generic' granule identifiers, distinguished by three things: ArchiveSet, ESDT, and Granule Key. Appending a timestamp can be used to find a particular granule.

`.../Dataset/17/OMTO3/2009-12-01T17:15:28`

A dataset includes all the granules of a given ESDT (e.g. OMTO3) that were part of the Archiveset (e.g. 17) at a particular date/time (e.g. 2009-12-01 at 17:15:28). These identifiers are suitable for inclusion in citations.

4.2 Artifact Web Server

The root of the PURLs (`http://purl.org/NET/ACPS/`) is redirected to a specific host, port and path that is responsible for resolving each of the artifact identifiers. Each one is either further redirected to some other server responsible for that artifact or handled by the ACPS server. For some identifiers, the actual bits making up the artifact are returned. For others we have defined a set of metadata for each artifact.

The server complies with the typical implementation of Representational State Transfer (REST [3]). It uses the `Content-Type` and `Accept` HTTP headers to determine the appropriate format. It can format and express that metadata in a variety of formats as desired by the requester based on the intended use of the information:

- **YAML**[3] *YAML Ain't Markup Language*: a very simple, human friendly format very useful for debugging and testing.

- **XML**[4] *Extensible Markup Language*: a standard for data transfer. We have (some) XML Schemas[5] that foster interchange of information with our system. It is also easy for others to parse the data with standard XML libraries and transform it in custom ways with XSL Transformations[6].

- **JSON**[7] *JavaScript Object Notation*: a lightweight data-interchange language that is particularly easy to incorporate into dynamic AJAX web site GUIs.

- **RDF/OWL**[8] *Web Ontology Language*: foundational formats for the Semantic Web[9]. Where appropriate we encode the metadata we associate with each artifact as OWL properties. This also includes the relationships each artifact has with other artifacts. This data can be ingested into off the shelf triple stores which can then support complex queries, reasoning and data mining.

Since the system distributes the data associated with each artifact, including links to other artifacts, it is easy to traverse the provenance hierarchy from any point in the greater graph to the specific depth required.

5 Semantic Web and Linked Data

While our focus so far has been on the internal functioning of the ACPS, we publish our provenance graphs to the Web, and by following specific standards and formats, the Semantic Web. We can include links to other entities on the Semantic Web and establish object equivalences and relationships with other entities following the principles of Linked Data. [10]

Quite often our data flows depend on other data sets (e.g. the ozone retrieval relies on an accurate snow and ice map dataset we acquire from the National Snow and Ice Data Center) other geophysical models often use our ozone data sets as inputs to their work. Each organization and data processing system should similarly publish their objects such that we can all point to one another in a much larger graph than that held by a single system.

6 Future Work

This system is still a work in progress, but has made strides in its organization and presentation of the artifacts of provenance for its own data flows. We are striving to remain compliant with the XML and RDF/OWL representations of provenance in the Open Provenance Model (OPM[11]) so that it will be interoperable with systems that can handle such provenance representations. We are also experimenting with Proof Markup Language (PML[12]) as a way to relate information with our provenance graphs.

In particular we are working to address the data citation issue. Users who do research with enormous Earth Science data sets that are constantly undergoing revision and reprocessing often have trouble conveying persistent and correct references to the specific data their research was based on. Our schemes for associating individual granules and their provenance to datasets and their summarized provenance can be useful in this area. Our work allows a researcher to use a single, persistent URI to refer to an entire data set under consideration. From that single citation URI, a reader can mechanically (i.e. not manually) work back through the provenance graph to determine all of the specific granules, executables, algorithms, calibration data, etc. that led to its existence.

A related issue is comparing two data citations to determine the difference. If two dataset references are not exactly identical, it can be burdensome to track down their difference. By traversing our provenance graph, we can determine the **relative equivalence** of two datasets. This can range from "equivalent" for two datasets that were produced independently, but using identical inputs under identical environments to "scientifically equivalent" where the data sets are close enough numerically that scientific studies using the data will produce identical results. Other more substantial differences can be determined where, for example, a calibration change was introduced or a major algorithm change.

7 Acknowledgments

The authors would like to acknowledge contributions from the ACPS software development team and numerous discussions with the NASA Earth Science Data Systems Working Group (ESDSWG[13]) and the Federation of Earth Science Information Partners (ESIP FED[14]) Data Preservation Cluster, in particular Ruth Duerr and Bruce Barkstrom.

References

[1] Chapman, A., and Jagadish, H. V. *Provenance and the Price of Identity*. Springer-Verlag, Berlin, Heidelberg, 2008, pp. 106–119.

[2] Clark, T., Martin, S., and Liefeld, T. Globally distributed object identification for biological knowledgebases. *Brief Bioinform 5*, 1 (2004), 59–70.

[3] Fielding, R. T., and Taylor, R. N. Principled design of the modern web architecture. *ACM Trans. Internet Technol. 2*, 2 (2002), 115–150.

[4] Jun Zhao, C. G., and Stevens, R. In *Provenance and Annotation of Data* (2006), vol. 4145 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 254–269.

[5] Tilmes, C., Linda, M., and Fleig, A. Atmospheric Composition Processing System (ACPS). *Geoscience and Remote Sensing, IEEE Transactions on 47*, 1 (Jan. 2009), 51–58.

Notes

[1] http://www.ietf.org/rfc/rfc2445.txt
[2] http://purl.org
[3] http://www.yaml.org/
[4] http://www.w3.org/XML/
[5] http://www.w3.org/XML/Schema
[6] http://www.w3.org/TR/xslt
[7] http://www.json.org/
[8] http://www.w3.org/TR/owl2-overview/
[9] http://semanticweb.org/
[10] http://www.w3.org/DesignIssues/LinkedData.html
[11] http://openprovenance.org/
[12] http://inference-web.org/
[13] http://esdswg.gsfc.nasa.gov/
[14] http://www.esipfed.org/