Earth Sci Inform (2010) 3:59–65
DOI 10.1007/s12145-010-0046-3

RESEARCH ARTICLE

Tracking provenance of earth science data

Curt Tilmes · Yelena Yesha · Milton Halem

Received: 30 September 2009 / Accepted: 24 March 2010 / Published online: 9 April 2010

Abstract Tremendous volumes of data have been captured, archived and analyzed. Sensors, algorithms and processing systems for transforming and analyzing the data are evolving over time. Web Portals and Services can create transient data sets on-demand. Data are transferred from organization to organization with additional transformations at every stage. Provenance in this context refers to the source of data and a record of the process that led to its current state. It encompasses the documentation of a variety of artifacts related to particular data. Provenance is important for understanding and using scientific datasets, and critical for independent confirmation of scientific results. Managing provenance throughout scientific data processing has gained interest lately and there are a variety of approaches. Large scale scientific datasets consisting of thousands to millions of individual data files and processes offer particular challenges. This paper uses the analogy of art history provenance to explore some of the concerns of applying provenance tracking to earth science data. It also illustrates some of the provenance issues with examples drawn from the Ozone Monitoring Instrument (OMI) Data Processing System (OMIDAPS) (Tilmes et al. 2004) run at NASA's Goddard Space Flight Center by the first author.

Keywords Data processing · Provenance

Communicated by: Thomas Narock

C. Tilmes (✉)
NASA Goddard Space Flight Center,
Greenbelt, MD 20771, USA
e-mail: Curt.Tilmes@nasa.gov

Ye. Yesha · M. Halem
University of Maryland,
Baltimore County 1000 Hilltop Circle,
Baltimore, MD 21250, USA

Ye. Yesha
e-mail: yeyesha@umbc.edu

M. Halem
e-mail: halem@umbc.edu

Provenance

The term and concept of provenance has a very long history (Moreau et al. 2008b). Perhaps its most familiar use is in art history. The provenance of a particular work of art is a critical component related to the authenticity, and ultimately, the worth of the work. If a painting, for example, is signed "Rembrandt" but there is no provenance related to the particular painting, no history of its creation and curacy during its lifespan a historian attempting to authenticate the painting will be very skeptical. If however, there are records accompanying the painting stating the date, place and circumstances surrounding its creation and an unbroken chain of custody backed up with verifiable records, deeds of sale, etc. the historian will have an easier time of authentication. This isn't to say that those records are prima facie evidence for authenticity—they are necessary, but not sufficient to make the case.

Similarly for earth science data, there has sometimes been a tendency to "sign" the data, stating where it came from, but neglecting to maintain the entire history of the data in a form suitable for verifying the "authenticity" of the data. For a single painting, the provenance

chain is generally pretty straight-forward. For an earth science dataset, it can be very complicated (Simmhan et al. 2005; Bose and Frew 2005).

Earth science provenance

There are several different classes of users for Earth Science Data. Most simply wish to use the data as inputs to their own research. They trust that the data served by an archive or distribution system is exactly what it says it is. Data producers and archives have a responsibility to ensure that the data they distribute is correctly identified. But some data are suitable for certain purposes, but not suitable for others (Suarez-Sola et al. 2008). Provenance information can help potential users of a given data set understand more about the data:

- *Where did a result come from?*
  Just as scientific research builds on research that has gone before, citing previous work as appropriate, earth science research results from the analysis of earth science data sets and should cite them. Data citations must be able to resolve the particular data that were used to produce a given result.
- *Who (organization/team/scientist) produced the result?*
  In an ideal world, concepts like "reputation" wouldn't have a place in science—results would stand or fall by their own merits regardless of their origin. In reality, reputation plays a key role in science, not in the absolute proof of correctness of a result, but in the establishment of the prima facie case for its correctness. Just as one might put more trust in art provenance stating that our painting was held in special collections at the Louvre in Paris for 100 years rather than Joe's Art Emporium, one might have more trust in data distributed by an earth science archive well known for its diligent curation.
- *Were the results from two independent analyses derived from the same data?*
  One of the big problems in earth science data is change. As science marches forward, we learn better ways to capture and calibrate data, better ways to retrieve geophysical parameters from the data and better ways to represent and format that data and package it up for end users.
  If two different researchers claim to have used temperature data from the TEMPSAT instrument but came to different conclusions, is the difference in their analyses or in their data? We can only answer that question with data citations to a degree of granularity sufficient to track their provenance and show that the data were indeed the same.
- *How can I independently reproduce the experiment to confirm (support) or refute the finding?*
  Again, there is a need to distinguish datasets to a fine enough degree of granularity that an independent researcher can obtain data that are sufficiently equivalent to reproduce the analysis under consideration.
- *How much should I trust the result?*
  Ultimately just as the provenance was so critical for the art historian to verify the authenticity of our painting, scientific data provenance contributes to the trust of a scientific result. If you can't point to the source of data used, and can't reproduce the results, you can't trust them.

Identifiers

Identity is a hard problem. Consider the identity of a person. Everyone has a name, but in practice even something as simple as a person's name is fraught with problems. Many people share the same name, or change their name at some point in their lives. Even if you have the correct name for a person, trying to resolve it can be a major burden—there are hundreds (thousands?) of potential directories purporting to list names of people. Some governments assign national IDs, like the U.S. Social Security Number, but that is considered private information, so can't be released. Email addresses are unique, but again, they change, and some people don't like releasing them where marketers can obtain them. The Friend of a Friend (FOAF)[1] uses a one way cryptographic hash of an email address. This seems to be a good compromise that protects the privacy of the address itself, and if reasonable efforts are made to protect them from reuse can maintain their usefulness as a unique identifier even if a person's primary email address changes.

One goal of representing provenance is to determine a unique, persistent, public identifier for every "artifact" involved or related to a particular scientific result. These include:

- **Data**
  Every data file must be identified. This includes unique identifiers for different versions of the data.

[1] http://www.foaf-project.org

- **Data sources, sensors/instruments/platforms**
  The original source of data must be identified. This could be an instrument on a satellite, an aircraft, a ship, or a graduate student walking around.
- **Test data/calibration data**
  Raw data from an instrument must be calibrated to determine the physical quantities that are being measured. Test data is used to characterize the instrument performance and determine the appropriate parameters to the calibration process.
- **Algorithms/source code/executables/documentation**
  Earth science data undergoes a myriad of transformations. From the raw data, through application of calibration, to geophysical retrievals to various gridding and formatting steps they are continually massaged by various algorithms. The algorithms are expressed in source code which are then compiled into executables which directly process the data. All of the above (if you are lucky) are documented. For future researchers trying to replicate those transformation steps, or maintain long term trends by applying similar techniques to data from future instruments, these are critical aspects of data provenance.
- **People/teams/projects/organizations/systems**
  As earth science data passes through the various processing steps, someone is responsible for that work. The identity of that entity should be recorded.
- **Computer systems/OS/compilers/libraries/formats**
  In general, the particulars of the environment surrounding the data processing shouldn't impact the results, but there have been occasions where they have. Representing this type of information with standard identifiers makes it easier to track down other data that could have been similarly affected.
  Constructing a complete operational environment capable of reproducing a given result isn't always trivial. Even if the precise components aren't available, providing a complete documentation of at least one such functional environment with a "known working" configuration will certainly aid in the efforts of anyone trying to reproduce a given result, or construct a similar environment in the future.
- **Validation data/analyses**
  After data have been processed, they are compared to other data to verify their correctness. This is analogous to an analysis of our unproven painting to known, verifiable examples. Though the analysis isn't, strictly speaking, part of the lineage or heritage of the painting, it is clearly documentation relevant to the establishment of the authenticity of the painting. Similarly, validation experiments are critical in establishing the correctness of earth science data and an important part of the provenance of that data.
- **Published, peer reviewed scientific papers**
  Again, these might not be part of the lineage of the data, they are still part of the body of knowledge contributing to the credibility of the results of various analyses. Many of the artifacts of earth science data processing are submitted as journal articles and subject to peer review. Such documents contribute greatly to the credibility of any scientific results.
- **Abstract events**
  Abstract things like a data transformation event or a validation experiment or even an ephemeral execution of a web service are perhaps the most difficult to come up reasonable identifiers, but that is the only way to document their relationship with all the other artifacts.
- **Multiple versions of all of the above**
  Rigorous configuration management must be used to properly and completely document provenance. Everything changes, and maintaining proper identity and documenting relationship during such change can be difficult, but this is the only way to answer the provenance questions above.

Each of these represent artifacts that have a relationship to the data under consideration. They should be assigned globally unique, actionable, persistent[2] identifiers, and identifier equivalences should be maintained across system or institutional boundaries so they all contribute to the global knowledge base. Good identifiers enable complete representation of the artifact relationships and also enable external annotations where others can contribute to the knowledge about given artifacts. There are various schemes for mapping identifiers into Uniform Resource Identifiers (URIs) that can be resolved through the World Wide Web.

Provenance representation

Capturing complete provenance information during data processing and archiving it is just the first step. For provenance information to be useful, it must be available to users of the data beyond the data producers themselves. Historically it has been represented in a hodge-podge of methods and formats, mostly as natural

[2] *Cool URIs don't change*, http://www.w3.org/Provider/Style/URI

language rather than any formal structured representation. A text description of the data source and processing chain can fill some of the needs, but is difficult for a machine agent to follow and must by necessity summarize information rather than scrupulously recording the details.

As the various artifacts undergo change (instrument re-calibration, updated algorithms, bugs fixed, etc), the relationships among the artifacts change as well and the provenance needs to be updated. The approach has usually been to discard the older versions of data, replacing them with the latest and best versions. Any provenance information stored as metadata associated with the files was generally discarded as well—even if the data are cited as inputs to some analysis in the published literature. Once the provenance information is lost, scientific reproducibility of previous results is virtually impossible (Heinis and Alonso 2008) . If provenance information is saved at all, it is often represented in non-standard forms that are difficult to follow. Imagine a phone call to a researcher asking "where did you get this data, and exactly what did you do to it?" It is very difficult to convey the precise information, and represent it in a form that can be understood by the receiver. Even if provenance information is kept around, some systems can't (or won't) reproduce older datasets. They sometimes rely on an error prone, manual process to attempt to reproduce data that were previously released.

Through a series of "Provenance Challenges" (Moreau et al. 2007) and workshops (Freire et al. 2008) the provenance community has been converging on a general model for provenance, *The Open Provenance Model* (OPM) (Moreau et al. 2008a). There are several proposed representations for provenance information that follow the model, including XML and XML/RDF. Another standard is the RDF/OWL based *Proof Markup Language* (da Silva et al. 2006) which includes (among other things) support for semantic tagging of the relationships between the artifacts of a provenance tree. We follow the general OPM and intend for our provenance representation to be interoperable with community standards as they emerge.

Provenance examples

Consider a basic example from the OMI Data Processing System (Tilmes et al. 2004), where some simple production rules select input files to feed to an Total Ozone data retrieval process (`OMTO3`):

1. Find the OMI Level 1B UV radiance file (`OML1BRUG`) for the corresponding orbit.

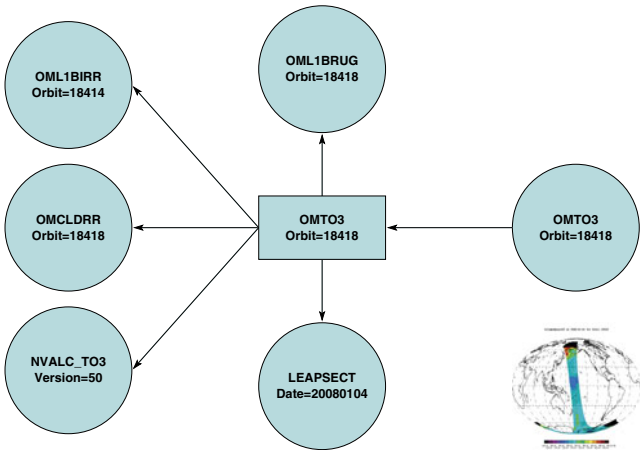

Fig. 1 Provenance example: input data files for a single granule or OMI data

2. Find the OMI Level 1B Solar irradiance file (`OML1BIRR`) from the nearest orbit. Irradiance measurements are made by OMI once per day, so the appropriate file may not be for the exact same orbit.
3. Find the OMI Level 2 Cloud file (`OMCLDRR`) for the corresponding orbit.
4. Find the latest lookup table of OMTO3 calibration values.
5. Find the latest Leap Second definition file.

Executing those production rules for process `OMTO3`, for orbit number `18418` could result in the processing flow shown in Fig. 1.[3]

Following the graph back a step to see where each of the input files themselves came from as shown in Fig. 2 we have a problem. One of the files was received from a developer, but we haven't fully captured the process by which it was created.

Figure 3 extends the provenance a little to capture not only input files, but also the processing environment (host `ominion607` was used to run the process. We also record as much information as possible about that host, including OS version, library versions, memory, processor, etc. Since that information changes over time (e.g. OS gets upgraded), it is tied back to the execution time of the process. We also record the version of the integrated algorithm that was used to perform the process. It is stored in the system, and distributed to end users in addition to the data files.

But consider that integrated algorithm, where did it come from? We have another process that was used

[3]The graph notation follows *The Open Provenance Model* (Moreau et al. 2008a), arrows point from artifacts back to inputs from which the artifacts were derived.

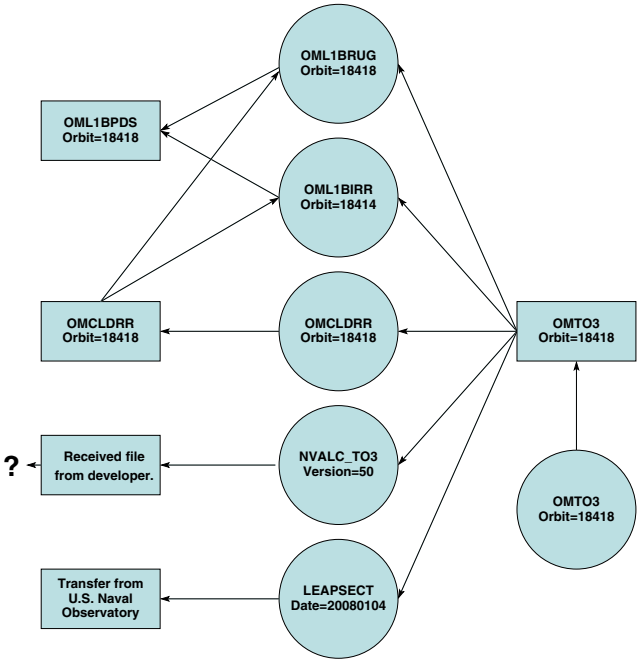

Fig. 2 Provenance example: missing link

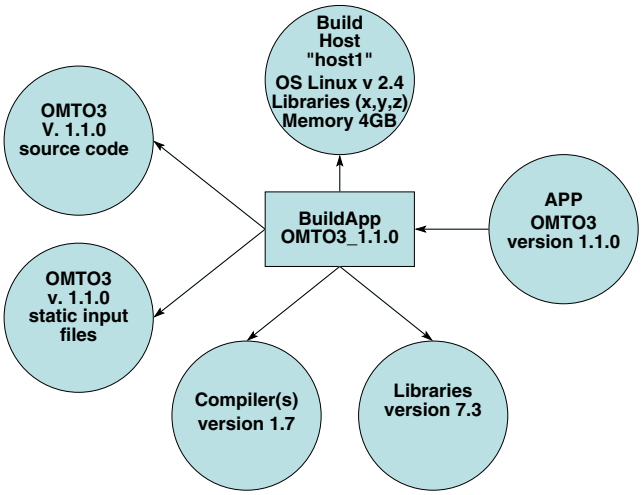

Fig. 4 Provenance example: include build event provenance to produce APP

to integrate the algorithm (that specific version of the algorithm) from some source code and various static input files as shown in Fig. 4. We also capture the build environment just like the processing environment, including what compiler versions were used, what library versions were used, etc.

Follow the provenance back a little bit more, where did the source code come from? It is documented in an Algorithm Theoretical Basis Document, backed up by research presented in a journal article as shown in Fig 5. Again, this isn't a strict lineage relationship, since such papers are generally published after the data have been processed, but the artifacts have a very strong relationship that is part of the overall provenance and should be documented for further understanding of the users of the data.

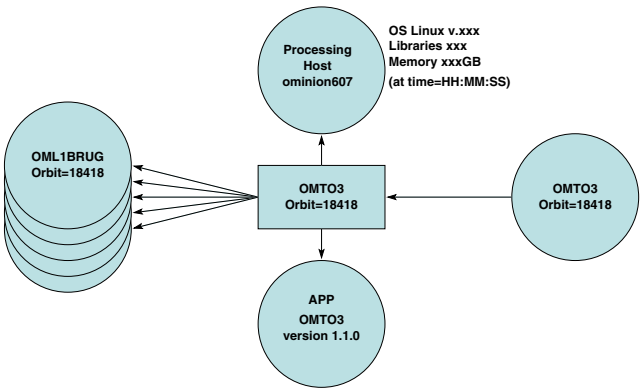

Fig. 3 Provenance example: include environment and algorithm

Reproducibility

One of the major goals of provenance is to enable reproducibility of results. Some artifacts are *essential* for reproducibility and some are not. From the list above, some artifacts that are considered essential include the data and the algorithms. Non-essential artifacts could be the person performing the processing, or the name of the computer where the processing took place.

Even though they aren't essential for reproducibility, they still provide a record of the process that led to the production of the data. That record provides an audit trail that lends credibility to the entire process.

A comparison of the provenance of two data sets evaluating their equivalence must consider all of the essential artifacts, but can disregard those artfacts that are non-essential.

Some artifacts, such as the version of a numerical library could be essential, but might not be. An investigation might be needed to consider whether or not the

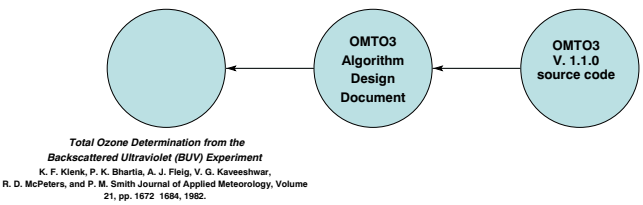

Fig. 5 Provenance example: link source code to design and published research

changes between the versions has a meaningful impact on the output.

Scientific reproducibility

While provenance information is nice to have for a researcher trying to understand a data set and algorithm, especially for climate research using remote sensing data, it is critical for scientific reproducibility. Many systems recognize this ideal and strive to store sufficient information that a dedicated (sometimes very dedicated) researcher who is able to expend sufficient effort could theoretically construct a system capable of reproducing the data. Other systems can reproduce the latest version of the data, but do not support obtaining older data.

Scientific Equivalence is not *Perfect Equivalence*, and *Scientific Reproducibility* is not *Perfect Reproducibility*. It is unrealistic to believe that an independent researcher can replicate the processing scenario perfectly. Our goal is to provide 1) sufficient information to reproduce the data close enough to come to the same scientific conclusion, and 2) describe the differences between two processing scenarios so that if there is a difference in the result, there is an avenue of analysis for determining the nature of the difference. In particular, all physical measurements of the real world are subject to accuracy and precision. As the data are processed and integrated with others, those metrics will change. The metrics will have varying sensitivity to the various algorithms and processing environment differences. Part of the commitment to doing science at this level involves developing robust algorithms that are less sensitive to changes in the processing environment (compilers, libraries, hardware, etc.) Within our own system, we make a concious effort to maintain the ability to reproduce older versions of data. We maintain unit tests and "golden" datasets with known correct answers, and as changes are introduced into our system, either algorithm, environment or hardware, we perform formal regression tests to evaluate the effects of those changes. This process can never be perfect, but we try to do the best we can.

Our goal is to make it not just possible, but *easy* to reproduce any data file that gets distributed from our system, both within our system (as described above), and for an independent researcher to confirm our results. To that end, we archive all versions of fully integrated algorithms. As a next step, we plan to distribute a processing framework that can access the integrated algorithms directly and interact with our system to download the information needed to replicate the environment and re-run the algorithms. This framework takes several forms depending on the requirements of the end user. As its most basic, for a user with compatible hardware (we use Linux/Intel) a single small script can follow the provenance record for a single granule, download the integrated algorithm and the required input files (if still online) and re-run the program to produce the data file. The next level up can process granularity iterators to reproduce a data set with many granules. The framework will also aid in exploration of the provenance graph and comparison of two provenance graphs.

Additionally, since the integrated algorithms are available and encapsulated provenance information is available, remote users could use their local execution framework to reproduce any of our files within their own systems. As "cloud" data processing systems such as Eucalyptus (Nurmi et al. 2009), Amazon EC2[4] and NASA's own Nebula[5] become more mature, we plan to build virtual compute environment images that can be used on those services. This will enable scientific reproducibility and allows an independent verification and validation of all data provided by our main system. Providing this capability can increase the *credibility* of the science results that use the data.

Conclusion and future work

Access to complete provenance information is essential for many aspects of the use of Earth science data. It is possible to build science data processing systems that automatically capture the provenance information with little impact on resources, operations or the participating scientists involved in creating the data. This will make it possible for users to know how a data set was made, reproduce the results of the initial processing, and understand differences over time periods even after the original producers are no longer available for consultation.

With complete provenance a user will know what input data was used for any product including details of where it came from and what version it was. The user will be able to know what exact algorithms were used to make a product, what exact input data was used, what exact system the data was produced with and what processing system it was made on. This will improve the credibility of the data set and make it possible to determine whether differences over time of a remotely sensed data set come from true geophysical changes or are artifacts of the production system.

[4] http://aws.amazon.com/ec2

[5] http://nebula.nasa.gov

We are working on development of methods to distribute the processing framework used to make a product in such a way that remote scientists can access the algorithms that were used, interact with our system to download the information needed to replicate the environment, and run time parameters and reproduce the results or modify any component and asses the impact of the change.

Complete provenance requires that input data obtained from external sources also comes with its own provenance. We are working on identifying tools, content, and standards for this and on encouraging other data sources to provide this information. We note with concern that there is not a current commitment in the science community to require adequate stewardship to maintain and support complete provenance even if it is available. We also note that the requirement for the scientists providing processing algorithms to also provide complete documentation of the final version of their process does not receive adequate support.

It is our hope that by showing that all of the needed information can be captured if available and that it can help enable scientific reproducibility we can encourage further development in these areas.

Acknowledgement Thanks to the NASA MODIS and OMI Data Processing teams.

References

Bose R, Frew J (2005) Lineage retrieval for scientific data processing: a survey. ACM Comput Surv 37(1):1–28. doi:10.1145/1057977.1057978

da Silva PP, McGuinness DL, Fikes R (2006) A proof markup language for Semantic Web services. Inf Syst 31(4–5):381–395. doi:10.1016/j.is.2005.02.003, http://www.sciencedirect.com/science/article/B6V0G-4FSCJS5-7/2/3c3bf9533f53cdacdc7cb2466e94e825 , the Semantic Web and Web Services

Freire J, Missier P, Moreau L, Schreiber A, Mattoso M, Silva CT (2008) Provenance and annotation of data and processes, vol 5272/2008. Springer, Berlin. doi:10.1007/978-3-540-89965-5

Heinis T, Alonso G (2008) Efficient lineage tracking for scientific workflows. In: SIGMOD '08: proceedings of the 2008 ACM SIGMOD international conference on management of data. ACM, New York, pp 1007–1018. doi:10.1145/1376616.1376716

Moreau L, Ludäscher B, Altintas I, Barga RS, Bowers S, Callahan S, Chin GJ, Clifford B, Cohen S, Cohen-Boulakia S, Davidson S, Deelman E, Digiampietri L, Foster J, Freire I, Frew J, Futrelle J, Gibson T, Gil Y, Goble C, Golbeck J, Groth P, Holland DA, Jiang S, Kim J, Koop D, Krenek A, McPhillips T, Mehta G, Miles S, Metzger D, Munroe S, Myers J, Plale B, Podhorszki N, Ratnakar V, Santos E, Scheidegger C, Schuchardt K, Seltzer M, Simmhan YL, Silva C, Slaughter P, Stephan E, Stevens R, Turi D, Vo H, Wilde M, Zhao J, Zhao Y (2007) Special issue: the first provenance challenge. Concurr Comput: Practice and Experience 20(5):409–418. doi:10.1002/cpe.1233

Moreau L, Freire J, Futrelle J, Mcgrath R, Myers J, Paulson P (2008a) The open provenance model: an overview. Provenance and annotation of data and processes, pp 323–326. doi:10.1007/978-3-540-89965-5_31

Moreau L, Groth P, Miles S, Vazquez-Salceda J, Ibbotson J, Jiang S, Munroe S, Rana O, Schreiber A, Tan V, Varga L (2008b) The provenance of electronic data. Commun ACM 51(4):52–58. doi:10.1145/1330311.1330323

Nurmi D, Wolski R, Grzegorczyk C, Obertelli G, Soman S, Youseff L, Zagorodnov D (2009) The eucalyptus open-source cloud-computing system. In: CCGRID '09: proceedings of the 2009 9th IEEE/ACM international symposium on cluster computing and the grid. IEEE Computer Society, Washington, DC, pp 124–131. doi:10.1109/CCGRID.2009.93

Simmhan YL, Plale B, Gannon D (2005) A survey of data provenance in e-science. SIGMOD Rec 34(3):31–36. doi:10.1145/1084805.1084812

Suarez-Sola I, Davey A, Hourcle JA (2008) What are we tracking ... and why? AGU Fall Meeting Abstracts, pp C1047+

Tilmes C, Linda M, Fleig A (2004) Development of two Science Investigator-led Processing Systems (SIPS) for NASA's Earth Observation System (EOS). In: Geoscience and remote sensing symposium, 2004. In: IGARSS '04. Proceedings. 2004 IEEE International, vol 3, pp 2190–2195. doi:10.1109/IGARSS.2004.1370795