IN11C-1047

What Are We Tracking ... and Why?

I. Suárez-Solá (NSO)
A.R. Davey (SAO)
J.A. Hourclé (NASA-GSFC/Wyle)

contact info : http://vso1.nascom.nasa.gov/docs/wiki/ContactUs

Virtual Solar Observatory
http://www.virtualsolar.org/

Abstract:

It is impossible to define what adequate provenance is without knowing who is asking the question. What determines sufficient provenance information is not a function of the data, but of the question being asked of it. Many of these questions are asked by people not affiliated with the mission and possibly from different disciplines. To plan for every conceivable question would require a significant burden on the data systems that are designed to answer the mission's science objectives.

Provenance is further complicated as each system might have a different definition of 'data set'. Is it the raw instrument results? Is it the result of numerical processing? Does it include the associated metadata? Does it include packaging? Depending on how a system defines 'data set', it may not be able to track provenance with sufficient granularity to ask the desired question, or we may end up with a complex web of relationships that significantly increases the system complexity.

System designers must also remember that data archives are not a closed system. We need mechanisms for tracking not only the provenance relationships between data objects and the systems that generate them, but also from journal articles back to the data that was used to support the research. Simply creating a mirror of the data used, as done in other scientific disciplines, is unrealistic for terabyte and petabyte scale data sets.

We present work by the Virtual Solar Observatory on the assignment of identifiers that could be used for tracking provenance and compare it to other proposed standards in the scientific and library science communities. We use the Solar Dynamics Observatory, STEREO and Hinode missions as examples where the concept of 'data set' breaks many systems for citing data.

What is Provenance?

Provenance can refer to different concepts with significantly different scopes:

Libraries and Museums :
History of ownership of an item ("chain of custody")
Archives :
The source of an item received (only previous owner)
Archaelogy (provenience) :
Where an item was found
Botany :
The location of the plant from where the samples were collected
Geology :
Where a rock was originally formed
Computer Science (data provenance) :
The pieces of data used to compute the item
Computer Science (process provenance) :
The execution history of computer processes used to compute the item

All of these definitions are related to tracking the source of a given item, but they track many different aspects of that origin:

The creation of the object
Origin of the materials used to create it
The processes that went into it
The person or group that created it
The handling of the object after creation
Ownership and control
Location and storage

Provenance In Solar Physics

As we perform observations, we must record information about the origin of the sensor recordings. A large amount of metadata is collected to allow scientists to properly make use of the sensor recordings:

Engineering metadata
operating mode of the sensor
environmental information (eg, sensor temperature)
Sensor location
Pointing information
Time of the observation

For data that is stored with calibration or other processing, the files typically contain additional metadata about the processing. For telescope data, many discipline scientists prefer to use the level 0 data and reprocess it themselves rather than worry about the potential process provenance of the higher level files. For data requiring more complex processing, such as magnetograms or helioseismology products, the level 1 files are the typical product used by the community.

Our granularity is the individual files, which are typically discrete images, but may be a data cube of files with similar observing modes or from a single sensor for a given time slice.

For many solar physics active archives, there is a single file served for a given sensor recording, but the file may change over time as newer calibrations are applied or metadata is updated. Other archives may maintain multiple files or different calibration levels, processing or file format, and either make assumptions about the best file to serve, or return multiple records for a given search.

Granule Identifiers

The Virtual Solar Observatory tracks data granules by a composite key, composed of the Data Provider identifier plus a provider supplied identifier ("file ID"). Although the file ID may be a file name, file path, or be parsed for other metadata, it is treated as an opaque string by the VSO, and passed back to the data provider to resolve for ordering. This allows providers to respond with the current calibration of the data, should a given granule be requested months or years later, perform negotiation to determine the best packaging, or to distribute the load across mirrors.

Unfortunately, as providers can give different responses for the same request, this means that we cannot rely on these identifiers to attempt to validate research by reproducting it, if these were the only identifiers used. Rots *et al* (2007) suggest that this is not unique to solar physics.

The identifiers are not used for de-duplications of the returned record sets, as the VSO makes no assumptions about how the identifiers are assigned by each provider, and we cannot identify which parts may be significant. They typically are assigned for use by the primary investigation, and may not be meaningful on their own for the greater scientific community.

"Shopping Cart" Identifiers

As people find data of interest, then can flag them to be placed into a "shopping cart" for later download. The data are tracked using the previously mentioned file IDs, which can be used to identify the general sensor recording, but not necessarily the exact state of calibration.

The shopping cart also tracks the queries that lead to the data being selected, nor the post-query filtering that the scientist may have applied. The intent of this process provenance is not to reproduce the output, but to allow scientists to see if the results change, as there may be additional data that was not available or was restricted when the query was originally run.

Provenance Use Cases

Provenance is important not just for the data, but for research that may make use of the data. Given a journal article, it is very difficult to do the following:

Identify the data used in the scientific paper.

Identify the processing performed for the author to come to their conclusion.

Determine the likelihood of the data having been mishandled by the researcher.

Determine the likelihood of the data having been corrupted before it got to the researcher.

Unfortunately, not all processing is done within a workflow system. A researcher may download the data and then subset it, sample or otherwise filter the data to reduce it to a managable size. They may then process the data to enhance the feature they are trying to observe, and to reduce what they believe to be background noise. The final data that they analyze and present may be very different than what was retrieved from the archive (Hanish, 2007). Unfortunately, these are details that may be necessary to validate a researcher's results.

Although citing a collection identifier, such as archive's notion of a 'data series' or a 'data set' or a user-defined collection such as the VSO 'shopping cart' may give others a starting point to reproduce the work, but we currently have no way to reliably document the full process. Although workflow systems may be able to generate much of the necessary documentation, it is unrealistic to believe that we can all researchers to use them, as they may have a number of proprietary tools, or may not have access to such systems.

The last use case would require archives to maintain records of not only the processing and data provenance, but also the storage provenance and would ideally require integrity validation through the use of checksums and mirrored copies.

Proposed Standards

Trustworthy Repositories Audit & Certification (CRL, 2008):
Requires checksums to be maintained in the archive, and the repository must document migrations and losses (A3.8), must validate the objects ingested (B1.4) and their origin (B1.3), as well as actively validating the integrity (B4.4).
The requirements that all objects must have a persistent, unique identifier (B2.5) creates additional complexities on active archives, but as it does not say that they must be the only identifiers used, active archives may still be able to comply with this requirement while continuing to serve an identifier to 'the most recent calibration'. TRAC does not help with data provenance once it has left the archive, nor with trying to identify larger collections in a relatively compact way.
Altman & King, 2007
The proposed requirements on data citation are a good start, but the issue of what a "data set" is for telescope data is tricky, and the date of publishing for a continuously updated collection is near impossible to determine. The idea of a numeric checksum is a further problem due to the sheer size of the collections we are dealing with. The idea of a "bridge service" to translate a short identifier is good, and to some degree can be handled by the 'VSO Cart' or similar service, but it would likely only link to either the archived or processed data, of which there might be millions of individual objects, and may not capture the processing provenance by the researcher. They suggest describing the subsetting performed in the text of the article, but this may leave us with the soft language currently in use.
Digital Object Identifiers (Paskin, 2005):
The use of persistant names, rather than storage location as with URLs is important, this mechanism only allows us to use an existing "bridge service" as described in Altman & King. The previous problems mentioned still remain.

Peer Reviewed Data Publication (Dallmeier-Tiessen and Pfeiffenberger, 2008):
By linking to an intermediary, peer-reviewed article that documents the dataset, we have advantages over just using OAI, as different data journals can place requirements on provenance documentation that may be required for their discipline. This still runs into issues with subsetting and other handling of the data once it has left the archive.
OAI Object Reuse and Exchange (OAI-ORE):
ORE provides for a way or citing compound digital objects. Although the aspect of alternative objects is important if we are truly trying to cite the data, and not the file containing the data, we leave that aspect for other discussion (see tomorrow's talk, Hourclé, IN22A-04). Where it is useful is in specifying aggregate objects, as we can create a file which specifically spells out the individual objects that were used in the research, and then provide an identifier to that single file. Unfortunately, this may require identifying millions of objects depending on the nature of the research and may only link back to the original archived objects, not the final processed data.

The Problem with "Data Set"

For time series data, it is easy to identify a continuous set of data. Unfortunately, with telescope data, we have a series of images that can be combined in any number of ways to form a "data set". As telescopes may run different compaigns, the observing mode, cadence, pointing, and other observing parameters are highly variable. Archives do not maintain identifiers for the various sets, as they are not built around that concept.

If someone refers to the set of 'SECCHI COR2' data, this is not only a problem that there are two discrete telescopes in entirely different locations (STEREO-Ahead and STEREO-Behind), but the telescopes alternate between polarization angles, which makes the issue of how the data was sampled quite significant for some types of science. The 'set' of COR2 data may be one, two or four (or another two) sets, depending on the researcher.

Hinode SOT can be subsetted by detector head (SP4D and FG), of which FG can be broken down by filter head (WB aka BFI vs. NB aka NFI), specific filter, or by observing mode (eg, 200ms exposure vs. 100ms). This results in 74 discrete datasets just grouping by filter and observing mode, but researchers may consider consistent cadence, pointing, or any other parameter to define their "data set".

References

Altman, M. and King. G. (2007). A Proposed Standard for the Scholarly Citation of Quantitative Data, *D-Lib Magazine* 13(3/4).

CRL (2008). *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist*. <http://www.crl.edu/PDF/trac.pdf>

Dallmeier-Tiessen, S. and Pfeiffenberger, H. (2008). Peer Reviewed Data Publication, *CODATA 2008*, 5-8 Oct 2008: Kyiv, Ukraine.

Hanish, R.J. (2007). Long-Term Preservation of Astronomical Research Results, *Science Archives in the 21st Century*, 25-26 April 2007: Adelpi, MD.

OAI. (2008). *Object Reuse and Exchange*. <http://openarchives.org/ore/

Paskin, N. (2005). Digital Object Identifiers for Scientific Data. *Data Science Journal*, 4:12-20. doi:10.2481/dsj.4.12

Rots, A., Accomazzi, A. and Eichhorn G. (2007). Associating Persistent Identifiers between Trustworthy Repositories, *Science Archives in the 21st Century*, 25-26 April 2007: Adelpi, MD.