Provenance and Volunteered Geographic Information

James Frew

Donald Bren School of Environmental Science and Management
University of California, Santa Barbara

Provenance (also called *lineage*) is metadata about an object's origin and history (Bose and Frew, 2005.) The term is conventionally applied to works of art, whose provenance is the documented chain of custody of the object, from creator to current custodian. Reliable provenance (assurance that the object has never been without reliable custody) is a necessary precondition for establishing an artwork's authenticity.

For information, the notion of provenance broadens to include the transformations applied between information's origins and its current form. (One can think of artwork provenance as special case where a change in custody equates to an identity transformation.) In my own work in data-intensive Earth science, a typical processing sequence includes: data acquisition by a satellite remote sensing system, formatting, calibration, projection, subsetting, re-projection, and analysis; the end product being a quantized field representing some Earth surface phenomenon such as snow cover or ocean color. The provenance of the end products is conceptually a directed acyclic graph leading back through transformations and intermediate data to the original satellite and ancillary data. Note that this graph can be traversed in either direction; i.e., one can determine both the "ancestors" and the "descendants" of any particular data object (e.g., file) or transformation instance (e.g., program invocation.)

The provenance of scientific information can be exploited to answer common, non-trivial questions. For example, *forward* (descendant) provenance can identify data products that were derived, however indirectly from suspect (e.g., miscalibrated) source data. *Backward* (ancestor) provenance can be used to help identify the source(s) of observed anomalies in a data product.

I assume that provenance will be useful, perhaps even critical, for the broader acceptance and utilization of volunteered geographic information (VGI), for three reasons. First, geographic information collected or manipulated by nonspecialists is more likely to contain unnoticed errors or biases. If and when these errors or biases are discovered, provenance can document (and thus help mitigate the consequences of) their propagation. Second, provenance can substitute for other missing or incorrect metadata, by identifying antecedent objects from which such metadata may be inherited. And finally, provenance can identify the humans or institutions involved in the information's creation and manipulation, and thus provide the foundations for judgments about quality and trust.

Supplying provenance manually is tedious, and like most human-created metadata, it is usually distinguished by its absence. A conspicuous, if partial, exception to this rule is, of course, article citations, which constitute a weak form of provenance in that it can usually be assumed that the references are all in some way ancestral to the referring document. The twin motivations of scientific integrity and professional advancement help ensure that this particular form of metadata is consistently supplied. The exception is partial, however, since the nature of the transformation is not explicitly specified; all we can state

reliably is that the cited documents somehow contributed to the citing document.

Aside from published documents, the overwhelming majority of scientific information (indeed, information in general) has little or no provenance associated with it. Most such information is created and processed in environments that do not capture and maintain provenance automatically, and there is no motivation, comparable to citation counts, for supplying provenance manually.

As part of my research in data-driven Earth science, I have developed a system that automatically captures the provenance of arbitrary computational sequences, and saves this metadata in a form such that arbitrary portions of the provenance graph can be easily retrieved and displayed (Frew at al, 2007.) The system has demonstrated interoperability (see http://twiki.ipaw.info/bin/view/Challenge/ES3_2) with other systems that maintain provenance information (e.g., workflow environments), so it is reasonable to expect that a standard will emerge for communicating and assembling provenance for distributed information; that is, information whose antecedent data and transformations span the Internet.

There is one form of metadata that *everyone* creates voluntarily: an HTML hyperlink, a fact exploited by web search engines. In particular, Google treats hyperlinks as implicit endorsements of the targets by the linking page, and ranks pages accordingly. This is strikingly similar to provenance, in that the endorsement is directional and the sum of the links forms a directed graph, but it's also quite different, since there is no way to tell (at least explicitly) whether a link denotes an ancestor-descendant relationship, and if so, which direction the relationship runs. (Also, web link graphs can easily contain cycles.)

I believe these twin technologies -- automatic metadata capture, and web hyperlinks -- can be combined to capture and maintain provenance for VGI. Locally, capturing metadata automatically is the only realistic way that such metadata will be made available: experts have enough trouble creating metadata; we cannot rely on volunteers to do so. On the web, one can imagine a simple "microformat" (see http://microformats.org) using the hyperlink "rel" and "rev" attributes to denote explicit ancestor-descendant relationships. The missing pieces are the tools that seamlessly integrate the uploading of captured provenance and the creation of typed links into VGI publishing environments. I welcome discussion of how this might be accomplished.

References

Bose, R. and Frew, J., 2005. Lineage retrieval for scientific data processing: a survey. ACM Computing Surveys, vol. 37, no. 1, pp. 1-28. http://dx.doi.org/10.1145/1057977.1057978

Frew, J., Metzger, D., Slaughter, P., 2007. Automatic capture and reconstruction of computational provenance. Concurrency and Computation: Practice and Experience (online). http://dx.doi.org/10.1002/cpe.1247

James Frew
Created: 2007-10-31
Last modified: 2007-12-07 10:54