

Earth System Science and Spatial Webs (DRAFT 2004-11-20)

James Frew

*Donald Bren School of Environmental Science and Management
University of California
Santa Barbara, CA 93106-5131*

frew@bren.ucsb.edu

Earth system science information

Earth science is increasingly multidisciplinary (i.e., Earth *system* science) and increasingly data-driven:

- Research questions such as elaborating the global carbon cycle require close cooperation between (at least) biologists, geologists, oceanographers, meteorologists, and information scientists.
- Key tools in addressing such questions are the spaceborne remote-sensing systems deployed during the last decade, which are now yielding terabytes per day of raw observations, data rates that were formerly the exclusive domain of supercomputer-hosted models.

Together these transformations are driving the *decentralization* of Earth science information systems:

- As the number of disciplines involved in a research problem increases, the likelihood of the appropriate investigators being co-located decreases.
- As the volume and complexity of data streams increase, the likelihood their being accommodated by a single system decreases.
- There is also the decreasing likelihood that a critical mass of investigators will co-located with a critical mass of data streams (e.g. satellite ground data systems).

"Investigator-led processing", as exemplified by the Federation of Earth Science Information Partners (www.esipfed.org), is the emerging paradigm for both creating and publishing Earth science data and information. That is, the scientists who develop the techniques and algorithms for transforming observations into measurements and measurements into models, are also responsible for generating and disseminating the resulting data and information products. In this new paradigm, Earth scientists are both consumers and producers of data products, assuming roles once reserved for large centralized data centers. Moreover, these distributed science data consumer-producers must be able to *federate* into (possibly *ad hoc*) product chains, where one scientist's product (e.g., snow cover maps) becomes another scientist's input (e.g., to a runoff forecast model).

Web fundamentals

What does the Web mean for this kind of distributed, federated Earth system science? Recall that, at its most fundamental level, the Web is:

- An object transfer protocol (HTTP), as implemented by a set of Internet servers;
 - i.e., *ubiquitous service*
 - i.e., a *distribution* mechanism
- An object naming scheme (URI), as supported by a set of HTTP servers.
 - i.e., *universal names*
 - i.e., a *federating* mechanism

Therefore, if:

- Earth science information providers individually obtain and disseminate their products through HTTP services;

- Earth science information providers collectively agree on how their products and services are to be referenced via URIs

then we have the simplest possible mapping of Earth system science onto the Web.

Unfortunately even this minimal level of interoperability is nowhere near ubiquitous. There is as yet no universal agreement on:

- formats in which Earth science objects can be encoded for transmission;
- services by which common transformations of these objects can be requested;
- nomenclature for referring to products and services

Formats (e.g. HDF) and services (e.g., DAP, WxS) are receiving by far the most current attention from the computing, Earth, and GI science communities, although since a distributed system must always be prepared to accommodate a multiplicity of either, "victory" in the standards process is less important here.

Naming, on the other hand, is the federating glue that holds a distributed system together. You can't retrieve an object or invoke a service if you don't know what it's called.

The "non-spatial" web has evolved three mechanisms to deal with name discovery:

- search engines (e.g., Google), which look for occurrences of words or phrases in an object's content;
- directories (e.g., dmoz), which categorize URIs according to some particular cataloging scheme;
- informal patterns (e.g., *www.companyname.com*, */~username*, etc.), which make it easier to guess an unknown URI.

Search engines are of course the most wildly successful of these strategies, but are conspicuously unsuitable for discovering spatial information:

- Most of the search qualifiers for spatial information are non-textual;
- Content-based indexing (and therefore searching) of spatial information is still largely a research problem;
- The most successful search engine (Google) bases its ranking on a particular kind of human-created metadata (the hypertext link) that (currently) has no analogue for spatial data.

Therefore I assume that (for the near-to-medium term, at least), directories and naming schemes will be the primary structures from which a spatial web is built.

Spatial webs and naming

By "spatial web" I mean "web of spatial information", not "spatialized web of information". The latter leads logically to a "Digital Earth" model: point at a model of the world and access the information associated with the indicated location. While the Digital Earth is compelling, I believe that universal *access* to spatial information is a more immediate (and by no means solved) impediment to distributed federated Earth science. A web of spatial information is a necessary precondition for a successful Digital Earth.

The critical questions facing a "spatial web" are therefore:

- How do we name objects in a spatial web?
- How do we discover these names?

In the interests of space I will limit the following to a discussion of the first question.

The traditional Earth science approach to systematic object naming is to develop project-specific nomenclatures based on unique properties of data or its processing, or both. These kinds of names are called "semantic identifiers" since they encode object-specific metadata. For example MODIS granules managed by the EOSDIS core system are given names that encode the granule's type, version, datetime of acquisition, and datetime of processing. In effect, such names are equivalent to a "title" in a nonspatial cataloging system.

Semantic identifiers have the desirable properties of:

- <>uniqueness within a specific domain
 - <>i.e., the comprise a namespace);
- ease of distributed generation

Withing the MODIS community, the semantics of a MODIS granule identifier reveal crucial information about the granule, and also happen to guarantee uniqueness within the community. Outside the MODIS community, MODIS granule identfiers are opaque. They are not globally unique, but they can be made so if extend with information that identifies that community.

Most of the work on distributed object naming has focused on persistence (e.g. DOI, DRI, PURL), and therefore assumes that semantic identifiers are evil, since the contexts in which the semantics are understood cannot necessarily be preserved along with an object. However, as [Kunze] points out, "persistence is purely a matter of service": all persistent identifier schemes require the persistence of a supporting service infrastructure.

I therefore suggest that the Earth science community federate around a loose hierarchy of naming schemes. The broader community would assign nodes in the hierarchy, leaving the node owners to assign the underlying identifiers as they see fit. For example, if the broader community agrees to use MODIS as a node (or, more specifically, a namespace identifier), then a partial URI of the form MODIS/granuleID becomes a universal granule name. Appended to an appropriate server URL, it becomes a means of accessing services defined for the object (for example, <http://server/MODIS/granuleID>). The is the idea behind MODster [Frew], in which the sole service provided is redirection.

Reliable naming is especially critical given that many of the objects in a spatial web may be huge (e.g. a single MODIS level 1 granule is several hundred megabytes. [Gray] notes that "Over the last 40 years telecom prices have fallen much more slowly than any other communication technology", making it ever cheaper to move processing to data rather than the other way round. Eventually all spatial data may well be exposed through high-level services that implement any reasonable custom processing, but until then, it will be very important locate the most accessible copy of a large object, or especially the copy that happens to reside at a site that also implements the required level of custom processing. This kind of redirection to an appropriate data/service combination is straightforward with a approach like MODster.

References

[Frew]

Frew, James and Gallagher, James (2002) MODster: Peer-to-Peer Sharing for Remote Sensing Standard Products. IGARSS '02 International Geoscience and Remote Sensing Symposium. IEEE Computer Society, Toronto, Canada.

[Gray]

Gray, Jim (2003) Distributed computing economics. Microsoft Research Technical Report MSR-TR-2003-24.

[Kunze]

Kunze, John (2003) The ARK persistent identifier scheme.