James Frew is an Associate Professor of Environmental Informatics in UCSB’s Bren School of Environmental Science & Management. He has worked in remote sensing, image processing, software architecture, massive distributed data systems, digital libraries, computational provenance, digital archives, and crowdsourced geographic information. His current research addresses the automatic generation of data citations, and the applicability of array database technology to environmental data management.

Perspective

In an era of fine-grained web services, it is useful to recall the comprehensive functionality of the traditional academic citation. With a single (well-formed) citation—for example[1]—we (ideally) obtain:

- **credit**: We acknowledge the contribution of other scholars, specifically their influence on the work in which the citation is embedded.
- **provenance**: The cited work’s age and venue help situate it in the broader scholarly landscape, and inform our judgement about its relevance and credibility.
- **access**: The information provided in the citation is sufficient for us to locate (and hopefully access) the cited work (or a copy thereof.)

All of these aspects of a citation are important for discovery. Although we tend to think of traditional citations as driving manual searches through the published literature, in fact services like Google Scholar use them quite effectively to automatically locate cited works, even when the citations are incomplete or ambiguous. Moreover, human-readable citations, as opposed to opaque persistent identifiers, allow the works containing them to retain their meaning independent of the presence (or proper functioning) of the PID infrastructure.

While the scholarly community has hundreds of years of experience with creating and using citations to documents, citations to data (including spatial data) are still in their infancy. Document citations exploit the relative simplicity of document organization hierarchies (monograph, book→chapter, journal→article, etc.) Data citation, on the other hand, wrestles with issues like granularity (e.g., spatiotemporal subsetting) and allocation of credit (e.g., for collaboratively-maintained datasets). Content standards for data citations do not yet address these complexities. My colleagues and I propose[1] that the most appropriate solution to this problem (and one that also exploits some key properties of the database systems in which much citable data currently resides) is for data providers to generate correct and complete citations for each data request. In my lab at UCSB, we have demonstrated the feasibility of this approach by extending the OPeNDAP spatial data
access web service to support automatic citation generation for each unique OPeNDAP request. So, as a discovery tool, what will “Google Data Scholar” look like?

Reference