Title not required

The wikification of geospatial metadata

Darren Hardy*

Bren School of Environmental Science & Management
University of California, Santa Barbara, USA
dhardy@bren.ucsb.edu
21 May 2010

Abstract. For decades, metadata has been the ever-present, cure-all solution to heterogeneous data integration and use. Yet, high-quality, ubiquitous metadata is extremely rare in practice. Current volunteered geographic information systems may provide insights on how the scientific community can produce and manage metadata for geospatial data infrastructures.

Keywords. Geospatial data interoperability, Scientific knowledge generation, VGI.

Introduction

Metadata is "data about data," intended to facilitate data discovery, integration, and use (or reuse). Practitioners often standardize metadata syntax and semantics, but adherence to metadata standards is extremely rare in distributed systems, especially large or globe ones; this is hereafter referred to as the "metadata problem." GIS usually assumes strongly-typed spatial data representations, and GIScience researchers have developed disambiguation methods (e.g., toponym resolution or fuzzy boundaries) for spatial data that do not comply with these structures. These complexities make metadata important for geospatial integration and use.

Volunteered geographic information (VGI) systems, however, successfully integrate heterogeneous data sources on a global scale without solving the metadata problem directly. VGI systems use "best effort" geotagging methods and representations to avoid the complexity of richer GIScience approaches to georeferencing. Moreover, the VGI notion of metadata, and its production and management, is different than in geospatial data infrastructures.

In this short paper, I discuss whether the *VGI approach to metadata production* would benefit the scientific community for its increasingly large geospatial data infrastructures. I briefly characterize the metadata problem, then introduce key aspects of the VGI approach.

Metadata in the wild

Scientific communities have collaborated on metadata standards and conventions, such as CF [1] and its predecessor COARDS [2]. In our study of earth science datasets published via the OPeNDAP protocol [3], we found that they do not accurately follow these conventions. In fact, only a minority of them *claim* their convention (as required), and even of those, only a fraction *accurately* adhere to their stated convention.

In practice, scientific data sharing varies by discipline. Ecologists, for example, take idiosyncratic approaches to data sharing and reuse, which depend on disciplinary knowledge and social factors [4]. This metadata problem forces scientists to use specialized knowledge and manual effort for data reuse.

* Submitted as a discussion paper for the *Workshop on the Role of Volunteer Geographic Information in Advancing Science*, at the *Sixth International Conference on Geographic Information Science* (GIScience 2010), Zurich, Switzerland, 14 September 2010.

Lessons from the VGI approach

VGI systems like Wikipedia may provide some lessons for metadata production and management in geospatial data infrastructures.

1. They use seemingly *trivial geotagging* methods. Thus, they avoid the complexity (and implications) of GIScience approaches to georeferencing. Typically, they use decimal degree geographic coordinates in an *assumed* datum (WGS84) and without enforcing numerical precision. For example, in Wikipedia, the geotag for UCSB is $(-119.84813^\circ, 34.41254^\circ)$, where $0.00001^\circ \approx 1$ m precision[1]. Flickr™uses a similar approach, but saves context for how users select location (e.g., a zoom level on an interactive map) [5].
2. They use seemingly *trivial metadata structures*. Wikipedia uses *Wikitext* [6], a lightweight markup language, for its article content and metadata. Flickr uses perhaps the simplest metadata structure of all: tags which are simply any word or phrase in an uncontrolled vocabulary.
3. They use *bots* extensively. Wikipedia has hundreds of semi-automated programs to perform a wide variety of editorial functions from removing vandalism to extracting metadata to suggesting work.
4. They *promote "refined"* and *flag "unrefined" or problematic* content. The community identifies content that is exemplary or meets certain quality standards, and promotes this content. They flag any content that needs further work which is then suggested to those looking for content to work on. They also flag content that is subject to controversy or "edit wars" [7].
5. They do *lazy, but rapid* integration, with most popular first. Mashups and other rapid prototyping use service-level, rather than data-level, integration. They focus on APIs and "cookbooks" or working examples, rather than formalized specifications. This approach enables rapid integration, but also lazy integration since they target specific uses with partial APIs.
6. They have *complete histories* of revisions. The complete context for changes is always available (and easily accessible) when issues arise or for tools to utilize.
7. They rely on *data mining*. They have tools that search through content looking for ways to improve their service, and they provide APIs for anyone who wishes to mine content programatically.

Conclusion

GIScience should consider the *VGI approach to metadata production and use*. This paper presents ideas that may complement those for how GIScience might integrate the increasingly voluminous *VGI data* into geospatial data infrastructures. The novelty and practicalities of VGI production may benefit the scientific community as they confront increasingly large scale data integration issues.

References

[1] S Hankin et al. NetCDF-CF-OPeNDAP: Standards for ocean data interoperability and object lessons for community data standards processes. In *OceanObs'09*, Venice, Italy, 21–25 Sep 2009.

[2] COARDS. Conventions for the standardization of NetCDF files. from http://ferret.wrc.noaa.gov/noaa_coop/coop_cdf_profile.html, 10 Feb 1995.

[3] D Hardy et al. Metadata in the wild: An empirical survey of OPeNDAP-accessible metadata and its implications for discovery. In *AGU Fall Meeting*, San Francisco, CA, 11–15 Dec 2006. AGU.

[4] A Zimmerman. Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *Int. J. Digital Libraries*, 7(1):5–16, 2007.

[5] P Jankowski. Discovering landmarks and movement patterns from photo postings. Invited talk, Department of Geography, University of California, Santa Barbara, 29 Apr 2010.

[6] C Sauer et al. WikiCreole: A common wiki markup. In *Intl Symposium on Wikis*, Montreal, Quebec, Canada, 21–25 Oct 2007.

[7] F B Viégas et al. Talk before you type: Coordination in Wikipedia. In *40th Hawaii Intl Conf on System Sciences*, 3–6 Jan 2007.

[1] This format is likely a result of ubiquitous GPS units that report coordinates in WGS84 and approx. 1m precision.