CHAPTER 7

Semantic Challenges for Volunteered Geographic Information

Andrea Ballatore
Center for Spatial Studies, University of California, Santa Barbara,
aballatore@spatial.ucsb.edu

Abstract

Vast swaths of geographic information are produced by non-professional contributors using online collaborative tools. To extract value from the data, creators and consumers alike need some degree of consensus about what the entities of their domain of interest are and how they are related. Traditional information communities, such as government agencies, universities, and corporations, have devised informal and formal mechanisms to reduce the misinterpretation of the data they rely on, curating vocabularies, standards, and, more recently, formal ontologies. Because of the decentralized, fragmented nature of peer production, semantic agreements are more difficult to establish and to document in volunteered geographic information (VGI), severely limiting the re-usability and, ultimately, the value of the data. This paper provides an overview of the semantic issues experienced in VGI, and what potential solutions are emerging from research in geo-semantics and in the Semantic Web. The paradigm of Linked Data is discussed as a promising route to handle the semantic fragmentation of VGI, reducing the friction between data producers and consumers.

Keywords

volunteered geographic information; data quality; geo-semantics; linked data

How to cite this book chapter:
Ballatore, A. 2016. Semantic Challenges for Volunteered Geographic Information. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 87–95. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.g. License: CC-BY 4.0.

Introduction

The production of geographic information was, until a decade ago, the exclusive territory of professional surveyors and cartographers, working for governments and private firms. The combination of increasingly powerful, cheap, portable, and interconnected computers has opened up unforeseen possibilities for data collection and sharing beyond professional circles, with already tangible effects (Dodge & Kitchin, 2013). These non-professional mappers and cartographers carry out their efforts on online platforms, producing digital artifacts, such as maps, geo-databases, and gazetteers. This process can be seen as a form of collective communication about some phenomenon of interest (e.g. tourist attractions, noise pollution, or animal behavior). The communication is mediated by machines through the encoding of knowledge from human minds into data and the decoding of data back to knowledge.

To be able to perform this process, the communities that produce volunteered geographic information (VGI) need to devise a shared conceptualization of the portion of the world they want to capture. Questions about what entities exist, what their attributes are, what relationships they have to each other, need answers with some degree of consensus. For a myriad of reasons, such a consensus is often hard to reach. The world and its constituents can be described according to many different, and equally valid, conceptualizations (Smith & Mark, 2001). This problem, rooted in human cognition and communication, is often called 'semantic heterogeneity', and is observable in the ubiquitous vagueness, synonymy, and polysemy in natural languages.

In this chapter, I provide an overview of the semantic challenges that VGI producers and consumers face when describing and interpreting their data. I cover this issue from the perspective of geo-semantics, the discipline at the intersection of geographic information science, computer science, and knowledge engineering. First, I cover relevant work in semantics in the context of geography. Subsequently, I focus on the specific context of VGI, and its peculiar challenges. As a case study, I consider OpenStreetMap and its community. Finally, from a more technological viewpoint, I discuss the emergent Linked Data ecosystem, assessing its promises for more transparent, participatory, and democratic geographic information commons.

Semantics and geographic information

Geography is pervasive in human experience and natural language. On a daily basis, we navigate in and communicate about the geographic world, referring to natural and man-made entities such as roads, cities, mountains, and rivers. Our intuitive understanding of such concepts conceals the complexity that is encountered when trying to encode them in a digital form. The term 'mountain', for example, has a common-sense meaning, but also possesses dozens of

local and specialized definitions around the world (Janowicz et al. 2013). When an information system needs to answer the question 'Where is Mount Everest?' there is no single, context-free, cross-cultural way to produce an answer that will satisfy all users. The same consideration can be applied to virtually all natural geographic features, whose boundaries are vague, seasonal, or gradual. Man-made features, while obviously exhibiting more intelligible and crisp organization, are not exempt from heterogeneity, and can be described, categorized, and aggregated through alternative and incompatible conceptualizations.

Geo-semantics, as a subfield of geographic information science, is concerned with providing theoretical and applied means to handle the variations in these concepts, with the purpose of facilitating the creation and processing of information in computationally tractable terms. Standardization of units of measurement, nomenclatures, and other aspects of the geographic domain is indeed an important way to reduce semantic friction, and has been successfully applied to many domains, such as the CORINE nomenclature for land use. However, as Janowicz et al. (2013) argue, geo-semantics is not about imposing standards for what we mean by 'mountain', but should be rather about providing ways to preserve and handle the local definitions across heterogeneous datasets, enabling precise translation mechanisms. These vague geographic concepts are hard to formalize, and their intrinsic cultural grounding makes them poor candidates for long-term universal standardization.

One avenue of research in geo-semantics focuses on ontology engineering in support of conceptual modeling in geographic contexts (Kuhn 2009). Unlike 'big-o Ontology', a branch of Western philosophy interested in the deep structure of reality and being, ontology engineering does not aim at assessing what actually exists in the world outside the human mind, but has the task of constraining the usage of terms in the data towards the meaning intended by their authors. The underlying intuition lies in the usage of formal semantics, such as first order or description logics, to provide machine-readable, less ambiguous descriptions of entities and their relationships, which can be used to support data sharing, integration, and constrained forms of reasoning. Insights from this arena include the formal clarification of *identity*, *rigidity*, *role, is-a*, and *part-of* relationships, which wreak havoc when misunderstood in complex information systems (Guarino 2009). This program bears similarities with traditional forms of Artificial Intelligence, with which it shares the formal approach, but differs substantially in that it lacks the ambition to model common-sense knowledge through logic, aiming for more realistic and pragmatic purposes, such as the handling the meanings of 'mountain' in the Himalayas and in Ireland.

VGI and meaning

When even well-funded scientific and corporate organizations struggle to handle semantic heterogeneity, it should come as no surprise that VGI contributors

and consumers encounter substantial semantic problems in their work. In online collaborative projects, tensions between alternative conceptualizations of the same portions of reality are common, as a quick exploration of Wikipedia's talk pages would reveal. If they want to create value, VGI contributors interested in cycling are forced to confront, sooner or later, what exactly they mean by 'cycle way', 'cycle lane', and 'road quality', and whether these concepts fit different national and regional contexts different to their own.

Taking a broad stance on the scope of VGI, its semantic structures vary from well-defined and curated geo-referenced datasets, such as GeoNames, to unstructured social media content and blogs. The former are geo-semantically explicit, having the purpose of covering the entire world systematically. By contrast, the latter contain large amounts of geographic information–mainly vague references to place names. Much VGI semantics lies between these two extremes, for example in the case of folksonomies and centralized tagging platforms, such as Tagzania, WikiMapia, and Flickr. Such semantic approaches consist usually of a combination of a top-down, centralized definition of a conceptualization by a small elite, and the emergent semantics of bottom-up, unrestrained tagging.

The most popular VGI project, OpenStreetMap (OSM), deserves separate treatment. This cartographic project is geographically explicit, and produces data substantially more complex than that of competing efforts such as WikiMapia. The main dataset of the project contains an uneven (but impressive) object-based description of the entire planet, including its roads, buildings, parks, forests, lakes, etc. The conceptualization underlying this data is a semi-structured folksonomy, documented on a wiki website,[1] permitting the creation of any new term deemed necessary by users. For example, the term *amenity=university* is used to tag universities. Rather than a fixed ontology, OSM's conceptualization is a transient, evolving product, open to modification and negotiation. The project experiences therefore a tension between the technical need for a stable conceptualization, and the desire of contributors to express their local knowledge without a top-down interpretation of their world being imposed upon them.

Because of its openness, OSM is an ideal resource to study the semantic dimension of crowdsourced cartography. Using a combination of media, including a wiki website, forums, mailing lists, and software tools, contributors negotiate the conceptualization that underpins the data they produce, often disagreeing (Ballatore 2014). By exploring this digital corpus, it is possible to probe the interconnected dimensions of the largely asynchronous negotiation performed by VGI contributors. Most of the observable negotiation in OSM revolves around ontology engineering, i.e. the extraction of an explicit conceptualization from tacit knowledge, but in an informal, online setting (Ballatore & Mooney 2015). The dimensions of this negotiation can be summarized as follows:

[1] http://wiki.openstreetmap.org.

Topology and mereology. A topology is needed to represent geographic entities, defining how entities can be connected or contiguous, grounded in a theory of boundaries, interiority/exteriority, and separation. Additionally, a mereology is necessary to encode complex spatial entities, specifying how the parts relate to wholes, for example in the case of large buildings.

Simplification and adaptation. Many domains, such as land cover, road classification, and traffic regulations, have been conceptualized in national and international contexts. However, such conceptualizations are often too complex for the scope of OSM and are filled with technical terminology. In these cases, contributors choose an appropriate subset of the conceptualization, and adapt it to suit their needs. For instance, in France, the European CORINE Land Cover nomenclature has been imported into OSM.

Universalism and localism. A fundamental tension arises between the desire to develop a universal conceptualization that will be applied all over the world, and the need to tap into the heterogeneous and local knowledge of contributors. Initially, contributors attempted an Anglo-centric universal conceptualization, and subsequently, facing an explosion of complexity and spatial variation, fragmented it into regional or national schemas. Notably, the classification of roads has been problematic since the inception of the project, even within the English-speaking world. Similarly, contributors struggle with the complexity of the national road legislation, resorting to translatable national schemas.

Problems of equivalence. The tension between universalism and localism results in problems of equivalence between languages, for example in the conceptualization of restaurants in different countries. As indicated by linguistic translation theory, contributors need to express local concepts that do not have a direct translation in English, such as concepts that depend on local practices, laws, and vocabularies (e.g. courthouse). Specific local entities are often described into more general English terms, losing potentially more precise local knowledge.

Contested definitions. To constrain the intrinsic vagueness of geographic terms, lexical definitions of terms provide an important normative tool to construct a shared conceptualization. As in other domains, lexical definitions can help constrain the intended usage of terms, specifying the necessary and sufficient conditions for their application. Unsurprisingly, conflicts frequently arise about the lexical definitions in OSM. Problems occur when definitions are underspecified, lacking necessary detail, and when they are overspecified, including irrelevant or confusing details. Definitional conflicts result in classification conflicts in the data, when contributors disagree on whether individuals fit a category or not (e.g. is a building a church, a chapel, or a cathedral?).

Conceptual granularity. Information can be expressed at different conceptual granularities, for example describing a geospatial entity as a generic 'tree' or as a 'Pinus roxburghii'. For this reason, contributors often disagree on the level of detail to be included in the conceptualization. In principle, infinite knowledge can be elicited about an entity from multiple perspectives, and the choice of what details should be included is arbitrary, and driven by the desired application. When a category is too generic, its usage is not constrained enough and different conceptualizations emerge. Overly specific categories also cause problems, as they often involve jargon, and are little used. The production of VGI oscillates between different levels of conceptual granularity, in a balancing exercise.

The promises of Linked Data

As I have argued so far, the online production of geographic information faces substantial semantic challenges. VGI communities rely on open tagging and other lightweight semantic approaches to describe their data, which result in frequent inconsistencies, ambiguity, and high terminological heterogeneity, hindering the re-usability and interpretability of the data. The semantic friction encountered in the production and consumption of such data is tackled through different top-down and bottom-up strategies to constrain the usage of terms (e.g. adoption of existing standards, lexical definitions, etc.). For GIScientists, these issues point to exciting research questions. How can we design conceptual models and technologies to support communication about the geographic world, reducing the gap between consumers and producers? How can emergent Web technologies be harnessed to help contributors express their ideas in a clearer, less ambiguous way, without imposing centralized conceptualizations? How can we support the expression and alignment of complex local definitions in intuitive ways?

A promising answer to these questions lies in the Linked Data paradigm (Kuhn et al. 2014). Emerging from 15 years of research in the Semantic Web, Linked Data proposes to express information in an inter-linked data space, built on a triple-based formalism that expresses any data as *subject-predicate-object* statements (e.g. *Dublin is_capital_of Ireland*, *European_Union is_a Political_entity*). The dominant technologies in this arena are RDF (a simple format to encode triples) and OWL (a logical language to define ontologies, i.e. formal specifications of conceptualizations). The triples are hosted in dedicated triple stores, which are able to index, store, process, and retrieve triples more efficiently than general-purpose database management systems. Unlike traditional datasets, linked entities must have unique Web identifiers (URIs) to enable humans and machines to navigate the data space to retrieve definitions and relations with other entities. To query the triples, SPARQL and its spatial extension GeoSPARQL are currently the most widespread choice, providing a standardized access mechanism, roughly analogous to Web APIs.

As a toy example, let us consider a scenario: a tourist wants to state that they took a picture of the Colosseum in Las Vegas, and simply tags an image file with the string 'Colosseum', which might refer to a Roman building in Rome, to its kitsch replica being photographed, or to an obscure board game. Using the linked data approach, existing entities that match the string in the open knowledge base DBpedia (dbp:Colosseum, dbp:The_Colosseum_at_Caesars_Palace, where 'dbp:' stands for http://dbpedia.org/resource/), can be suggested to the tourist, who can then select the appropriate entity. The photo can now be described as triples (e.g. photo_001 is_a Photograph; photo_001 represents dbp:The_Colosseum_at_Caesars_Palace), which can be stored and processed automatically, inferring for example that the Colosseum is a theatre designed by the firm Scéno Plus Inc., enabling new avenues for data exploration and reducing the potential misinterpretation of the picture.

This simple idea has proved fruitful in both academic and industrial contexts (Heath & Bizer 2011). Notably, several Linked Open Data (LOD) initiatives have generated an ever-expanding cloud of interconnected datasets containing billions of triples.[2] VGI is a central pillar of this 'online commons' of re-usable open resources, providing the geographic ground for the organization of knowledge across domains. Projects like GeoNames, LinkedGeoData, and GeoWordNet form a constellation of open geo-knowledge bases (Ballatore et al. 2013). Major corporate actors such as Google and Yahoo! have also embraced Linked Data principles, offering increasingly structured search products based on RDF knowledge bases.[3] Media groups including the BBC and the New York Times publish part of their informational assets as Linked Data. Adopting a more lightweight, simpler approach, Microformats promote the semantic annotation of people, places, products, reviews, and organizations in Web pages, supporting the interpretation of content, without requiring the adoption of more complex Semantic Web infrastructure.[4]

Conclusions

In this chapter, I have summarized the challenges faced by VGI from a semantic perspective. First, I discussed the conceptual difficulties intrinsic to the vagueness of many geographic concepts. Second, OpenStreetMap (OSM) was taken as a case study to highlight the semantic issues that cause friction in the process of VGI production and consumption. VGI contributors coordinate their efforts and express information using a variety of semantic approaches, ranging from open tagging to controlled taxonomies and vocabularies. To produce intelligible data, OSM contributors make choices concerning topology and mereology,

[2] http://lod-cloud.net.
[3] https://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html
[4] http://microformats.org.

in a tension between universalism and localism. Particularly in transnational contexts, the description of entities encounters problems of equivalence, resulting in contested definitions of geographic concepts, at different granularities. To what degree these aspects extend to other VGI projects is an open research question. As a promising way to support the expression of heterogeneous local knowledge in VGI, I have briefly discussed Linked Data, a technical paradigm that has grown from Semantic Web research. Linked Data aims at constructing a data space in which geographic entities can be defined and described through standardized and precise logical mechanisms.

While I have presented reasons to support the adoption of Linked Data in VGI, the open challenges and current limitations of the approach cannot be ignored. The logic formalisms used in Linked Data, such as OWL, are rather ill-suited for spatio-temporal reasoning and need substantial extensions. The triple model, while conceptually attractive, can be very verbose to describe traditional geographic data such as raster images, and its structural complexity can explode quickly in realistic scenarios. To explore the current limitations of Linked Data, it suffices to take a closer look at the LOD Cloud, whose datasets vary hugely in their interpretability and noise, and whose interlinking is often patchy and uneven.

The approach promotes semantic clarity but cannot enforce it. Without formal constraints, Linked Data can be as obscure and ambiguous as plain text: the halo of clarity fades out as soon as poorly structured datasets are subject to integration and complex processing. Finally, the tools available to produce and process Linked Data often lack usability, and substantial design efforts are needed for deployment in VGI contexts, providing intuitive approaches to encoding local knowledge and alternative truths. None of these issues are insurmountable, and they do not outweigh the enormous potential benefits. Ultimately, the Linked Data paradigm should be considered as a promising technical framework to mitigate semantic problems in data production and consumption, and not as an unlikely fix to ancestral flaws in human communication that are here to stay.

References

Ballatore, A. 2014. Defacing the map: Cartographic vandalism in the digital commons. *The Cartographic Journal, 51*(3): 214–224.

Ballatore, A., & Mooney, P. 2015. Conceptualising the geographic world: The dimensions of negotiation in crowdsourced cartography. *International Journal of Geographical Information Science, 29*(12): 2310–2327.

Ballatore, A., Wilson, D. C., & Bertolotto, M. 2013. A Survey of Volunteered Open Geo-Knowledge Bases in the Semantic Web. In Pasi, G., Bordogna, G., & Jain, L. C. (Eds.), *Quality Issues in the Management of Web Information*. Berlin: Springer, pp. 93–120.

Dodge, M., & Kitchin, R. 2013. Mapping Experience: Crowdsourced Cartography. *Environment and Planning A*, *45*(1): 19–36.

Heath, T., & Bizer, C. 2011. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology, 1*(1): 1–136.

Guarino, N. 2009. The Ontological Level: Revisiting 30 Years of Knowledge Representation. In: Borgida, A., Chaudhri, V., Giorgini, P., & Yu, E. (Eds.), *Conceptual Modelling: Foundations and Applications. Essays in Honor of John Mylopoulos*. Berlin: Springer, pp. 52–67.

Janowicz, K., Scheider, S., & Adams, B. 2013. A Geo-Semantics Flyby. In: Rudolph, S., Gottlob, G., Horrocks, I., & van Harmelen, F. (Eds.), *Reasoning Web. Semantic Technologies for Intelligent Data Access* . Berlin: Springer, pp. 230–250.

Kuhn, W. 2009. Semantic Engineering. In: Navratil, G. (Ed.), *Research Trends in Geographic Information Science*. Berlin: Springer, pp. 63–76.

Kuhn, W., Kauppinen, T., & Janowicz, K. 2014. Linked Data–A Paradigm Shift for Geographic Information Science. In: *Geographic Information Science*. Berlin: Springer, pp. 173–186.

Smith, B., & Mark, D. M. 2001. Geographical categories: An ontological investigation. *International Journal of Geographical Information Science,* 15(7): 591–612.