CHAPTER 14

Querying VGI by semantic enrichment

Rob Lemmens*, Gilles Falquet†, Stefano De Sabbata‡, Bin Jiang§ and Benedicte Bucher¶

*University of Twente, The Netherlands, r.l.g.lemmens@utwente.nl
†University of Geneva, Switzerland, Gilles.Falquet@unige.ch
‡University of Leicester, UK, s.desabbata@le.ac.uk
§University of Gävle, Sweden, bin.jiang@hig.se
¶IGN, France, benedicte.bucher@ign.fr

Abstract

Volunteered geographic information (VGI) plays an increasing role in current geodata provision. At the same time, due to its lack of structure, it is hard to use as meaningful input in software applications. In this chapter, we embark upon the unstructured character of VGI and on ways to enrich the structure in order to make it suitable for information retrieval. We describe the characteristics of semantic enrichment and explain how folksonomies and ontologies play a role. We believe that they represent different levels of formality in a semantic reference space and determine the richness of the information retrieval.

Keywords

VGI, Query, Semantic enrichment, Folksonomy, Ontology

How to cite this book chapter:
Lemmens, R, Falquet, G, De Sabbata, S, Jiang, B and Bucher, B. 2016. Querying VGI by semantic enrichment. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 185–194. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.n. License: CC-BY 4.0.

Introduction

Recent developments in personal computing, GPS and Web 2.0 technologies are enabling a wide web audience to actively contribute to geo-information through the internet. Information obtained in this way – commonly referred to as volunteered geographic information (VGI) – is often difficult to query due to several reasons.

The complexity of querying is rooted in the informal, unstructured, heterogeneous nature of VGI, which is often published without a description of its context. Those issues are inherent to the process by which VGI is produced, i.e. by individuals who are in most cases not concerned with the query process. This chapter investigates how the process of querying VGI can be improved by semantically enriching it during its production and after it is published. The enrichment connects VGI to well-known concepts which are captured in both informal structures (folksonomies) and formal structures (ontologies). A folksonomy represents a particular domain through a set of user-generated tags/topics of domain-related information, whereas an ontology constitutes a domain more rigorously through the representations of logical relationships between concepts used in that domain. In this research we differentiate the semantic enrichment along the line of informal-formal conceptualization, i.e. evaluating conceptual bases ranging between folksonomy and ontology, supporting the enrichment of VGI.

The main point we want to stress is that VGI implies further degrees of freedom and expression for the users, which can enable new, different narratives in collecting, describing, and representing geographic information. At the same time, this intrinsic diversity requires the creation of 'interfaces' between VGI datasets and any algorithm aiming to analyze them, in order to translate the folksonomy (representing the vocabulary used in the VGI) into the structure used by a query algorithm. This is a challenge in terms of 1) the ad-hoc work necessary to deal with the data and 2) the errors, misinterpretation, and information loss in the translation.

We pose the following main research question and set the scene for its discussion, but do not claim to answer it yet fully: how does varying the level at which a top-down ontology is applied to a bottom-up folksonomy change the understanding of underlying data, and thus the ability of querying VGI?

The goal is to query VGI sources such as Tweets, commented photos and news items about the named features they contain. Typical queries are

- what is the location of a feature named X?
- what is the footprint of a feature named X?
- what are the features located at or near P?
- what are the features with type T?

In some cases this involves the harvesting of implicit geographic information (see also Kessler et al. 2009) and in other cases such information cannot be

directly extracted from the VGI itself and it needs to enriched with more formally structured information obtained from related sources, such as Wikipedia, OpenStreetMap, Geonames, etc. (see Smart et al. 2010).

Terminology of semi-structured data

VGI may appear as structured, semi-structured and unstructured data. Kitchin (2014) defines structured data as data 'that can be easily organized, stored and transferred in a defined data model', thus encompassing all data that can be represented and dealt with using relational databases and other technologies or representational models such as object-oriented languages or description logics. As a result, this kind of data can be straightforwardly processed through algorithms and visualized using graphs and maps. By contrast, semi-structured data don't have a, rigid, regular, or complete predefined data model/schema as required by traditional databases (Abiteboul 1997), though having 'a reasonably consistent set of fields' (Kitchin 2014). These include content that could barely be coded in a relational database, while still being characterized by irregular and flexible structures. Abiteboul (1997) provides a clear explanation of how HTML pages are a good example of semi-structured data, due to their lack of uniformity, and ample use of plain text. Finally, data is defined as unstructured if it has no structure that can be identifiable as common for the whole dataset, despite each element of the same dataset might have its own internal structure, which is not shared by any other element.

On such basis, most VGI content would be classified as semi-structured data, as few datasets could be straightforwardly dealt with in a relational database. Instead VGI datasets commonly use loose data definitions and categorizations, which are flexible and constantly edited by the same users, as well as more suitable to describe large quantities of vague information. A good example of loose categorization of geographic data can be found amongst the OpenStreetMap (OSM) map features — which include over eight thousand different user-defined kind of shops.

Most VGI content would also fit in Kitchin's (2014) definition of 'captured' data', that is data that has been directly captured through some device with the specific intention of capture the data. However, it might be argued that geotagged information, such as photos, entail 'exhaust' geographic data (implicitly included geodata) in the form of GPS coordinates in the image header—that is, as byproduct of capturing the photo, but not as main outcome of the process.

Characterizing the heterogeneity of VGI

VGI is very heterogeneous and diverse, due to three major reasons.

First, geographic features are very heterogeneous, since there are far more small geographic features than large ones. Using a more scientific terminology, geographic features are fractal or scaling (Jiang & Yin 2014), and they are best characterized by some heavy tailed (Zipf 1949) rather than Gaussian-like distributions. There are, for example, far more small street blocks than large ones. The small street blocks can be named as city blocks in cities, while the big street blocks are called field blocks in the countryside. The small street blocks constitute cities or natural cities to be more precise, whereas the large street blocks collectively form the countryside.

The heterogeneity of OSM data can be examined from various aspects such as element sizes, the number of edits, and the number of users for each element (Ma et al. 2015). For example, the element size ranges from 3 up to 5,000,000, the number of edits for each element can go up from 1 to 2,000. It is the heterogeneity that makes VGI unique and powerful in comparison to authoritative geographic information. It is the heterogeneity that makes VGI differ fundamentally from small data. It is the heterogeneity that makes researching VGI interesting and exciting. We should go beyond small data thinking such as Gaussian distributions and Euclidean geometry, and towards big data thinking such as heavy tailed distributions and fractal geometry.

Second, VGI can be produced through different methods and technologies, implying different levels of structural rigidity, ranging from menu entries to free text entries. This has important implications for semantic querying (see Section 6). The same VGI dataset may contain structured, semi-structured and unstructured data. For instance, the geometric part of OSM or the data/time metadata of Twitter are structured, as there is a fixed schema for them; additional information about geographic features in OSM consists of semi-structured sets of tag-value pairs; and the content of some fields are unstructured texts. In OSM, tags can be freely created by the user and the way people assign a geometry to a feature is not always consistent throughout the project (with different scales typically (Touya & Brando 2013)).

Third, VGI contributors may come from very diverse geographic, cultural, and technical backgrounds, and thus might be accustomed with different terminologies, or have different narratives. Some VGI is produced with a shared conceptualization that can be a set of tags or a category graph (like OSM or DBpedia), yet the production of data with this conceptualization in mind is done differently depending on contributors (Brando & Bucher 2010). An example is mapping of crimes, where people can interpret the levels of violence differently. Besides, sets of tags evolve over time. Hence, if data have not been tagged with a specific tag, it might just be due to the fact that that tag did not exist when they were produced.

The heterogeneity of geographic features, modes of production, and contributors' background are all contributing to the fact that the quality of VGI is often disputed and that even the quality itself is heterogeneous.

Folksonomies and Ontologies for querying VGI

Writing an algorithm to perform a task on a given data source, or querying this source, can be better accomplished if the meaning of each element of the source is well defined. In traditional structured sources this meaning is conveyed by a database schema or a datatype definition expressed in a database or programming language. In VGI the situation is different because 1) the schema, if it exists, may not be sufficient, due to various interpretations by the users, and 2) many VGI sources are only semi-structured, without any centrally defined schema. Therefore it is necessary to rely on some semantic resource to represent the meanings of the data elements.

There are several types of such semantic resources, ranging from the most informal (folksonomies or glossaries) to the most formal (formal logical ontologies). These resources, generally known as knowledge organization systems, can be characterized along two axes: 1) the structure complexity of the underlying data (tags, classes, hierarchical relations, etc.) and 2) the formalism used to express concept definitions. Figure 1 presents a classification, along these axes, of the most frequently used knowledge organization systems.

Semantic enrichment of VGI

Semantic enrichment refers to the process of making information more meaningful by adding explicit structure, metadata, definitions, etc. Explicit means that the result is queryable.

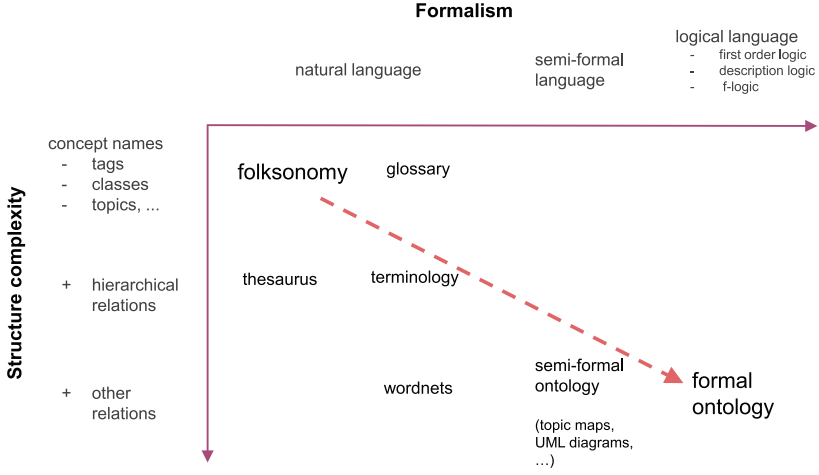

Figure 1: Informal and formal semantic reference space.

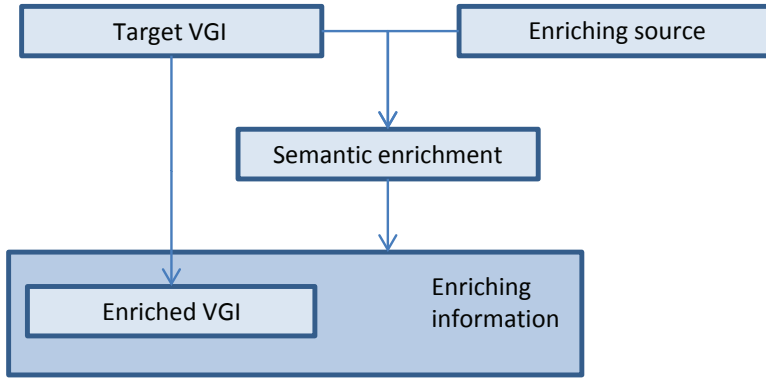

Figure 2: Semantic enrichment process.

The result of the enriching process obviously relies on the syntax and semantics of the target information and the enriching information source and the way in which the enrichment is performed (see Figure 2). We highlight three aspects which are crucial for a successful usage of the enriching process: 1) The semantics of the enriching source, 2) The semantics of the enriching information and 3) The syntax of the enriching information.

Targets can be free text in which case grammar rules provide the enriching information source and semantic enrichment is done through natural language processing (NLP) (see for example Peñas & Hovy 2010). In some cases the enriching information is intrinsically held in the target itself, such as relationships between items in a photo databases, such as Flickr. In such cases co-occurrence and data mining methods can be used to make these knowledge explicit (Deng et al. 2009).

In other cases the enriching information source is constituted by other sources, such as

1) (for exhaust-like data): web resources, sensors, gazetteer information (see Graham & De Sabbata 2015), etc.
2) (for 'captured data'): shared 'guidelines', each capturer's skills and intention, tasks assignment between several capturers, and the capturers' abilities to work together.

The second category can be captured by context models and provenance (background on how the information was produced) as reported in (Abel et al. 2012). The enriching information appears itself in different forms, for example as an ontology (see Lacasta et al. 2012).

In geospatial applications the semantic aspects of space put an extra constraint on semantic enrichment. Ballatore et al. (2011) combine a semantically-rich and spatially-poor ontology (DBpedia) with a spatially-rich and

semantically-poor VGI dataset (OpenStreetMap) to facilitate spatial knowledge discovery. As geo-information is so often constructed through multi-function workflows, provenance plays an important role in understanding the created geo-information. In addition, in VGI projects we think it is relevant to capture what people intended to do with the VGI at hand.

The result of the semantic enrichment can range from an ontology to more light-weight schema elements. Such enriching information can exist as separate entities relating to the target information or can be embedded within the target as metadata. As such they provide a more meaningful view on the target data and the basis for more meaningful queries, as described in the next section.

Towards semantic queries

In this section we show the different uses of semantic enrichment when querying a VGI source. For each structure level (structured, semi-structured, unstructured) we study what can be done with and without semantic enrichment: what are the problems and limitations with 'direct' queries on the VGI source and how different types of enrichment can help.

The main problem that arises when querying the structured part of VGI lies certainly in the differences in terms of quality and semantics that occur in the attribute values. E.g. a time value may be expressed in local time or in UTC, a length with different units of measurement, etc. These variations may ultimately render query results very imprecise or even meaningless.

The semantic enrichment of structured data may connect structural elements (table and attribute names) or data elements (attribute values) to semantic entities in some knowledge organization system.

The semantic enrichment of structural elements can be exploited by meta-level queries that help build correct queries. For instance: Find the tables and attributes that hold information about employment rates. This is particularly useful when the database schema is large and complex. Any type of knowledge organization system can be used for this purpose. In a geographical context, enrichment can be done for example by making geographic properties explicit, e.g. that a bridge is part of a road, 'built-up area' is an aggregate of building features, etc.

If the enrichment is done with a formal logical ontology, it becomes possible to express deductive queries (such as in the programming language Datalog) that can produce results not computable with standard SQL queries.

The semantic enrichment at the data level consists in associating attribute values to descriptors that make them meaningful (units, scale, accuracy, etc). These descriptions can then be used to augment the queries with selection criteria or transformation functions to produce higher quality results (e.g. select only those data that have a sufficient accuracy and a given unit of measurement).

Since VGI sources essentially link data to geographic entities, a natural enrichment consists in annotating the data elements to entities in some geographic knowledge source, such as Geonames or a geographic ontology. This will allow for semantic queries that combine geographic knowledge and other data.

Semi-structured VGI presents additional types of problems. Since the schema is generally not controlled, users can create multiple structures to represent the same real world phenomenon. For instance, the DBpedia database has at least five different properties to describe a person's birthplace, even though these names are obtained from supposedly structured 'infoboxes' of Wikipedia. This leads to complex queries in which all the possible property and value names must appear, e.g. {?x ex:placeOfBirth ?p} union {?x ex:birthplace ?p} union {?x birthPlace ?p} in a SPARQL query. The problem is of course worse with sources in which users regularly define new attribute names and values, as is the case in OpenStreetMap. In some situations it may become almost impossible to express consistent and complete queries.

For querying semi-structured VGI, the role of the semantic enrichment is essentially to describe and unify (or differentiate) the multiple naming schemes produced by the users. If the names used in the VGI source are associated to corresponding entities in an ontology, then the ontology's vocabulary can be used to express 'unified' queries that can be automatically rewritten into the VGI's vocabulary to produce a (complex) query on the VGI source.

In the case of OpenStreetMap, many tags may designate the same concept and a single tag may designate different concepts in different contexts. For instance, the semantic query roadType *motorway* will return features (roads) tagged with roadCategory motorway, roadCategory highway, roadCategory turnpike, category, turnpike, type motorway, etc. And in case the formal language models more relationship types, it will also indicate related features such as bridges, traffic lights, etc.

The abstraction level of the ontology used for the enrichment will determine the level of (semantic) detail of the queries. A high level ontology will unify many different names of the VGI into a single high level concept, while more precise (domain specific) ontologies will enable queries that are closer to the VGI's level of granularity. A similar remark applies to the geographic axis. The geographic precision of the results will depend on the scale and precision the geographic ontology that is used to enrich the VGI source. Moreover, if the enrichment structure possesses a rich semantic structure, with subclass relations or more sophisticated axioms, it will support a more expressive query rewriting. For instance, a query about Artists could be rewritten as a query about its subclasses Painters, Sculptors, Musicians, etc. if the Artist concept is not directly represented in the VGI source. In more than one case one should combine several ontologies to support the queries (see for example Lemmens & Kessler 2014).

In semantic enrichment a basic effort consists of associating texts with the concepts they deal with, this is generally accomplished with techniques such as word sense disambiguation and named entity recognition (in particular

geographic entity recognition). In this effort, one has to implement methods that are able to deal with the vagueness of information. With this kind of enrichment, semantic queries can answer questions such as 'find the data elements (texts) about the concept X'. Higher levels of enrichment consist in extracting precise information (facts) from texts. This amounts to transform unstructured sources into (semi-)structured ones, which is still an open research challenge.

Conclusions and recommendations

The semi-structured nature of VGI causes without doubt problems in the querying of its contents. We have presented several ways of imposing structure in order to facilitate more meaningful queries. Even the most basic queries, which go beyond text search, rely on some kind of structure. Whether the right degree of structure can be created depends on the success of the semantic enrichment process. In case of VGI, there are a variety of options, for which some of them rely on the reference to geodatabases. Semantic enrichment is basically constituted by linking the VGI to ontological concepts and their relationships.

We believe that some of the enriching information sources need curation as they are often ambiguous themselves. The level of enrichment needed, for which the semantic reference space is positioned between folksonomy and formal ontology, depends on the type of queries and needs to be further investigated with practical use cases.

References

Abel, F., Hauff, C., Houben, G., Tao, K., & Stronkman, R. 2012. Twitcident: fighting fire with information from social web streams. In: WWW 2012 Companion, pp. 305–308. DOI: http://doi.org/10.1145/2187980.2188035

Abiteboul, S. 1997. Querying Semi-Structured Data. In: *Proceedings of the 6th International Conference on Database Theory*. London, UK, Springer-Verlag, pp. 1–18. Retrieved from: http://dl.acm.org/citation.cfm?id=645502.656103.

Ballatore, A., & Bertolotto, M. 2011. Semantically enriching VGI in support of implicit feedback analysis. In: *Lecture Notes in Computer Science* (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 6574 LNCS, pp. 78–93. DOI: http://doi.org/10.1007/978-3-642-19173-2_8

Brando, C., & Bucher, B. 2010. Quality in User Generated Spatial Content. In: *A matter of specifications, 13th International Conference on Geographic Information Science (AGILE'10)*, 10–14 May, Guimarães, Portugal.

Deng, D.-P., Chuang, T.-R., & Lemmens, R. 2009. Conceptualization of place via spatial clustering and co-occurrence analysis. In: *Proceedings of the 2009*

International Workshop on Location Based Social Networks, pp. 49–56. DOI: http://doi.org/10.1145/1629890.1629902

Graham, M., & De Sabbata, S. 2015. Mapping Information Wealth and Poverty: The Geography of Gazetteers. *Environment and Planning A*: Forthcoming. Available at SSRN: http://ssrn.com/abstract=2587746.

Jiang, B., & Yin, J. 2014. Ht-index for quantifying the fractal or scaling structure of geographic features. *Annals of the Association of American Geographers, 104*(3): pp. 530–541.

Kessler, C., Janowicz, K., & Bishr, M. 2009. An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Seattle, WA: ACM. Retrieved from: http://dl.acm.org/citation.cfm?id=1653771.

Kitchin, R. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, London.

Lacasta, J., Nogueras-Iso, J., Falquet, G., Teller, J., & Zarazaga-Soria, F. J. 2013. Design and evaluation of a semantic enrichment process for bibliographic databases. *Data and Knowledge Engineering, 88*: 94–107. DOI: http://doi.org/10.1016/j.datak.2013.10.001

Lemmens, R., & Kessler, C. 2014. Geo-Information Visualizations of Linked Data. In: Huerta, Schade, & Granell (Eds.) *Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science*. Castellón, June, 3–6, 2014. ISBN: 978-90-816960-4-3.

Ma, D., Sandberg, M., & Jiang ,B. 2005. Characterizing the heterogeneity of the OpenStreetMap data and community. *ISPRS International Journal of Geo-Information, 4*(2): 535–550.

Peñas, A., & Hovy, E. 2010. Semantic enrichment of text with background knowledge. In: *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, (June), 15–23. Retrieved from: http://dl.acm.org/citation.cfm?id=1866775.1866778.

Smart, P. D., Jones, C. B., & Twaroch, F. A. 2010. Multi-Source Toponym Data Integration and Mediation for a Meta-Gazetteer Service. In: *Proceedings of GIScience 2010*. Zurich, Switzerland. Lecture Notes In Computer Science 6292, pp. 234–248.

Touya, G., & Brando, C. 2013. Detecting Level-of-Detail Inconsistencies in Volunteered Geographic Information Data Sets. *Cartographica: The International Journal for Geographic Information and Geovisualization, v48*(2): 134–143.

Zipf, G. K. 1949. *Human Behavior and the Principles of Least Effort*. Addison Wesley: Cambridge, MA.