Specifications for User Generated Spatial Content

Carmen Brando, Bénédicte Bucher, Nathalie Abadie

Université Paris-Est – Institut Géographique National (IGN)-COGIT Laboratory, Saint-Mandé, France
{carmen.brando-escobar, benedicte.bucher, nathalie-f.abadie}@ign.fr

Abstract. This paper addresses the issue of quality in the context of collaborative edition of spatial content. The overall approach is grounded on the definition of explicit and adequate specifications for such content, i.e. the data model, the conceptual model, conventions for data acquisition, possible integrity constraints, possible relationships with external reference data. Explicit specifications could be processed to automatically check when different users simultaneously contribute on the same area. Their definition requires expertness, firstly, to ensure spatial content consistency and, secondly, to establish relevant relationships with external reference data. Designing these specifications is not an easy task for contributors. Hence, the focus of this paper is to assist them in this task. We propose a generic process to automatically produce specification items such as feature types, attribute types, and relationship types, including possible relationship types with external reference data from a set of keywords. It exploits information from two different kinds of existing contents: user generated content (like Wikipedia) and more conventional content (like WordNet and NMA databases). It has been applied to keywords found in existing user generated spatial contents.

1 Introduction

The growth of user generated content (UGC) on the Web has lead to voluminous sources of information like Wikipedia. This trend applies to spa-

S.C.M. Geertman et al. (eds.), *Advancing Geoinformation Science for a Changing World*,
Lecture Notes in Geoinformation and Cartography 1, DOI 10.1007/978-3-642-19789-5_24,

tial content as well. User Generated Spatial Content (UGSC) stems both from the geotagging of existing UGC, such as Wikipedia articles, and from the edition of geographic features as is done in GeoNames, Wikimapia, or OpenStreetMap (OSM). This kind of content is known within the Geographic Information Science (GISc) community as Volunteered Geographic Information (VGI) (Goodchild 2007). The research community, governmental organizations, and businesses are more and more interested in using UGSC. There may be several motivations to use such data. It is free geographical data, a source for valuable update alerts for mapping organizations, and a source of complementary data which cannot be found in National Mapping Agencies' (NMAs) data sets.

An important stake for usability of UGSC is enhancing its quality. There are several challenges with respect to UGSC: challenges inherited from the "user generated" facet and challenges inherited from the "spatial" aspect. In particular, a major consideration to manage quality of conventional spatial content is to have an explicit structure for the content, e.g. classes and attributes, and conventions that rule unambiguous content acquisition (e.g. the road geometry is acquired at the middle of the road). This information has been designated geographic data set specifications (Abadie 2009). Specifications must be defined carefully to facilitate data consistency. They must be explicit for the user to know how (and how much) the data represent reality. Let us consider someone interested in withdrawing money; for doing so, he/she uses an application to look for ATMs (or cash machines). If he/she does not see ATMs on the map around his location, whereas there are ATMs in nearby areas of the map; he might infer that there is actually no ATM in his neighborhood, and decides to drive to another area in town. Lineage metadata could have been processed by the application to find out that contributors who have edited the map around his location have used a previous version of specifications where ATMs were not already included. In this way, the application could have informed the user that no ATM on this part of the map did not mean there were no ATMs in reality and then, he/she would not have taken the wrong choice. More generally, the relationship between geographical data and the reality cannot be fully assessed without knowing conventions and conceptual models that have ruled data acquisition. Last, user generated spatial content specifications should be formal, at least the semantics part in order to facilitate UGSC integration with other geographic information sources (Kavouras and Kokla 2008). In the remainder of the paper, the term model will refer to conceptual model and data model (ISO 2005).

This research work was carried out within the context of a PhD thesis, which aims at proposing novel methods for improving quality and usability of collaborative geographic data. Our approach is grounded on the im-

provement of specifications (i.e. enhancement and formalization) for this kind of content. Building such specifications is not an easy task for contributors; they possibly do not know enough about GISc to define an adequate model of classes, attributes, integrity constraints, and conventions. Thus, the present paper focuses on our proposal to assist them in this task. We conceive a process to automatically define a set of reusable modeling elements (i.e. feature, attribute, relationship types) dedicated to structuring UGSC. These elements consider relevant clues from two worlds: the world of user generated (spatial) content and the world of conventional geographic databases. The world of UG(S)C consists of large volumes of data available on the Web and of communities of contributors. The relevance of these sources with respect to our concern is that they are crowdsourced. Hence, they make use of non-specialized vocabulary well-known to contributors. The world of conventional geographic databases is typically led by private and public mapping agencies. Relevant hints from this world firstly are their techniques, especially with respect to the modeling of geometry and integrity constraints. Another important hint from conventional geographic databases is data themselves, which have undergone a specific quality checking process and have well documented quality metadata.

The remainder of the paper is organized as follows: section 2 explains briefly the relevance of content structure for quality management in UGSC and then analyzes existing propositions to structure UGC, conventional spatial content, and UGSC; section 3 details the process of building a set of feature types, attributes types, and relationship types for UGSC using several sources of information. It also includes implementation details and some results of this process using an example. Section 4 concludes the paper by recalling its main contributions and by announcing future work.

2 UG(S)C specifications

2.1 Relevance of specifications for UGSC quality management

Several aspects of quality management may be identified in UG(S)C projects (Brando and Bucher 2010; Antoniou et al 2010). Most of them refer to what can be called content specifications.

A first aspect is *internal consistency*. Several components in UGC enhance internal consistency. In wiki-powered sites, the word concerning a relevant concept is an HTTP link to the corresponding page. Whenever

these words are used in a page, the reader can follow the internal link (or wiki link) and obtain the definition of the concept. Some "semantics" may be added to the site by creating categories of articles, which help to reduce possible ambiguities. Internal consistency is also ensured by specific mechanisms to reconcile concurrent editions (Oster et al. 2006). With respect to geographic content, internal consistency also means not having conflicts between geographic features in the database, e.g. a house overlapping a road is usually a topologic conflict. Most of these conflicts are detected by the evaluation of integrity constraints which involve performing spatial operations on data. In other words, management of internal consistency can be enhanced with an adequate structure and explicit integrity constraints, which are part of the content's specifications.

Another aspect is the use of *references to external source*s. Wikipedia contributors are asked to quote external sources. In the world of geographic information, this may designate referencing objects in the real world based on an identifier attribute, for instance. It can also designate a reference to another data set (e.g. a NMA's data set).

A third aspect is a*uthority and reliability of contributors*. In the Google Encyclopedia Knol[1], quality management is mainly based on authors' identification and qualification. In standard metadata for geographic information proposed by ISO, this aspect of geographical data, namely its origin, is described in specific quality metadata: lineage information (ISO 2003). Managing authority and reliability of UGSC contributors has been addressed by Bishr and Kuhn (2007). Abilities of contributors have been empirically analyzed by Budhathoki et al. (2010) in the case of OSM. The authors suggest that those who take part in this open map-making are not laypeople as claimed in recent mainstream GIS literature; most of them have some prior experience in geospatial technology; and they are highly concerned about producing accurate and detailed maps. This seems to suggest contributors are aware of and care about existing specifications, at least in the case of OSM. A fourth aspect is c*omparison with a reference content* whose quality is supposed to be ensured, e.g. comparison between Wikipedia and Encyclopedia Britannica (Gilles 2005). This aspect has been extensively investigated by Haklay (2010) and Girres and Touya (2010). Such a comparison relies on data matching. Having formal specifications should facilitate this matching process because they include formal definitions of the meaning of classes and attributes which can be processed automatically to match data schemas priori to data instances (Kavouras and Kokla 2008; Abadie 2009).

[1] http://knol.google.com

At last, an important aspect of spatial content is the homogeneity of the acquisition process. The space represented by the content must be covered homogeneously. In other words, there should be no loss of balance of geographic space induced by acquisition biases, e.g. in UCrime[2], people are allowed to map criminal activities. Depending on the availability of contributors and their witnessing an assault, an area may be empty of crimes whereas there have actually been more assaults with respect to other neighborhoods. For OSM, Haklay (2010) has also observed heterogeneities in the description of geographic space. Specifications act as unambiguous guidelines for acquisition, hence facilitating the acquisition of homogeneous data. They can also be used to document data, hence to explain heterogeneities related to different versions of specifications used to cover different areas of the map.

This subsection has briefly exposed the relevance of specifications for UGSC quality management. The next subsections present an analysis of existing models to structure UGC, conventional spatial content, and UGSC.

2.2 Models to structure UGC

The best example of UGC is Wikipedia. There are certain elements to structure information within the encyclopedia. An article page explores a single issue and is mainly composed of a title, which summarizes the information concerning the issue in a phrase. It also contains content which discusses the issue in detail.

Furthermore, categories have been defined to annotate articles and organize Wikipedia content. There are many categories related to geography, notably physical geography, which contains sub-categories such as bodies of water, physical infrastructure, landforms, and natural disasters. A category page usually contains a list of the subcategories and articles referencing that category. It may sometimes include a brief description of the category. For example, "the dam category includes articles on dams in general. It includes man-made dams for flood control, hydroelectric power generation, transport, or water supply, as well as natural dams." Articles belonging to the same category may sometimes use a dedicated structure for summarizing information; it has been called an infobox. An infobox is a set of subject-attribute-value triples presenting some common aspects shared by several articles (Wu and Weld 2008). For instance, articles on individual London tube lines include the TfL (or Transport for London)

[2] http://ucrime.com

line infobox, which contains attributes concerning physical characteristics, statistical, and historic information. Similarly to categories, contributors tend to use infoboxes as a way of categorizing articles (Nastase et al. 2010). For instance, many articles concerning the World's mountains use the mountain infobox, which refers to a category.

Specific internal links allow setting up interesting mechanisms for improving the coherence of the entire encyclopedia. Firstly, disambiguation pages are meant to clarify the sense of a certain term (Mihalcea 2007). For instance, the term "plant" possesses several connotations; thus a disambiguation page titled *Plant_(Disambiguation)* has been added. It may refer to “living organisms” or “facility's infrastructures.” Secondly, redirection pages list alternative names for a single issue. For example, body of water and waterbody both have the same signification. Thus, waterbody is actually a redirect page which links to the page body of water. Lastly, Wikipedia is a multilingual encyclopedia which covers more than 25 languages. Every Wikipedia language edition is maintained separately. Pages would usually contain links to the corresponding pages in other languages. For instance, the page for category lakes *(*i.e. *Category:Lakes)* contains a link to the French version *(*i.e. *Catégorie:Lac)*.

An important community effort to extract structured information from Wikipedia is DBpedia[3], which is a knowledge base consisting of over one billion pieces of information from several language editions of Wikipedia. These elements are consistent with a cross-domain ontology, i.e. the DBpedia ontology, which has been manually derived from Wikipedia. DBpedia knowledge base covers general domains of information such as places, persons, organizations, species, etc. However, the coverage of every domain is not exhaustive. It may seem quite superficial for specialized areas of knowledge (e.g. Geography). Another issue is the availability of particular DBpedia data sets in other languages different from English. The most interesting resources (i.e. DBpedia and Infobox ontologies) are only available in that language. They do provide raw data sets in RDF triplet form for infoboxes, article's titles and abstracts, images' description, and internal links in almost any other language.

2.3 Models to structure conventional spatial content

Existing proposals to facilitate the design of geographic conceptual models (ISO 2005; Bédard et al. 2004; Parent et al. 1998) altogether highlight the relevance of feature types, attribute types, relationship types, geometry

3 http://dbpedia.org/

types, level of detail, and temporal aspects. A feature type represents a physical or abstract concept (e.g. road or land lot), and it usually has attribute types defined (e.g. a country's population). A relationship type allows establishing a connection between feature types.

Some of the usual relationships used in GI are composition and specialization relationships. An example of a composition is a feature type individual property which can be composed of a main building and a backyard. Another relevant relationship exclusive to geographic information is the relationship between features that represent the same object but at different levels of details. For instance, it may relate a representation of a city as a point and another representation with a polygon geometry (more detailed). It may also relate this representation of a city as a features collection: a set of buildings that make up the city. Other important relationships in GI are related to topology, distance, and orientation (Bruns and Egenhofer 1996). Most of these relationships are not explicitly specified but can be calculated by performing spatial operations on features' geometries (e.g. containment of districts within cities). Preserving these spatial relationships has always been a matter of concern during evaluation of spatial content consistency. Using a model with shared geometry is a strategy to preserve topological relationships. Another strategy is to use spatial integrity constraints (Mäs 2007). For instance, an integrity constraint indicating that administrative boundary lines are usually placed throughout the middle of waterways can be defined to improve spatial representation of the content.

Besides the definition of a conceptual model, conventional geographic data producers provide a documentation that explicitly describes using natural language how to encode the reality through the conceptual model. They ensure homogeneous capture of content especially when data collectors are different (Abadie 2009). They also help users to understand content (i.e. what to expect from it).

2.4 Models to structure UGSC

Even though part of Wikipedia is spatial content, we distinguish UGSC as content exclusively concerned with the spatial domain. Contributors of UGSC are usually acknowledged as neogeographers in the VGI world, i.e. people who have no academic or professional background in GISc but who are learning through practice (Turner 2006). They play a large role in ordering and categorizing spatial content (Graham 2010). UGSC projects encourage users to use a common vocabulary when editing content, typically by annotating geographic features by means of a user friendly GUI. These annotations are called categories in Wikimapia, tags in OSM, and feature

types in Google Map Maker (GMM). All these annotations will be referred to as tags in the rest of the paper. Examples of tags are historic buildings or state routes. Tags' meaning is documented in the help pages of UGSC projects. For instance, OSM provides all permitted tag values in its wiki pages[4]. Users are encouraged to use already defined tags, though they can freely define their own tags. OSM tags are classified in physical tags for material features, such as highways and waterways, non physical tags for abstract features, such as routes and boundaries, and naming tags for identifying features, such as common and official names of places. GMM distinguishes four main themes, natural features, roads, cities – political regions, and points of interest (POIs). Categorization schemes for both projects seem very exhaustive. For instance, not only pedestrian trail and wetland have been defined, but also less common POIs such as research centers and lighthouses. Wikimapia does not define any themes; all categories are at the same level. At least, this is not clearly available in the documentation pages.

2.5 Summary

To summarize, models to structure UGC, UGSC, and conventional spatial content have their advantages and disadvantages. They can all contribute to the creation of an adequate model to structure UGSC and facilitate its quality management.

UG(S)C consists of large volumes of data available on the Web. This content is crowdsourced by communities of contributors. UGC has been organized through two interesting mechanisms, categorization schemes and internal links. These elements help to enhance internal consistency of the entire content. UGSC tags represent an invaluable source of information about UGSC due to its volunteered nature. They represent a non-specialized vocabulary well-known to contributors, and more comprehensible for neogeographers than the usual argot used in NMAs' databases. OSM model is extensible because it allows contributors to define new tags. UGSC is meant to non-expert contributors but a certain expertise is required to understand the contribution process. Documentation of UGSC models is not quite exhaustive considering that UGSC tags can sometimes be ambiguous. For instance, the difference between mini-roundabouts and roundabouts may be difficult to establish for contributors.

A major consideration for conventional spatial content is to have an explicit structure for such content. It includes feature, attribute, and relation-

[4] http://wiki.openstreetmap.org/wiki/Map_Features

ship types. Examples of relevant relationships and properties for spatial content are composition or topological relationships and the level of details. It also includes conventions and integrity constraints, which rule unambiguous content acquisition. These elements help to enhance homogeneity during acquisition by several operators. Documentation of the content's model plays an important role to solve problems related to ambiguity of certain terms of the model (e.g. for feature types). Another important hint from conventional geographic databases is data themselves, which have undergone a specific quality checking process and their quality is well documented.

3 Proposal: building a predefined model for UGSC

Our proposal aims at facilitating the design of models for UGSC by acquiring predefined modeling elements from diverse sources of information, i.e. feature, attribute, and relationship types. More precisely, we have designed and implemented a process to automatically build modeling elements for UGSC from a set of user keywords. These elements include relationships with external reference data of which quality is known. To illustrate this process, we present an example based on UGSC tags extracted from the main French UGSC projects. The proposed process is also meant to be used on-the-fly to generate new user-defined modeling elements by specifying keywords. This section depicts the process and ends with a discussion about encountered difficulties and some clues to solve them.

3.1 Feature types and attributes types for UGSC

Wikipedia seems a valuable source of information for feature types (categories) and attributes types (infoboxes). Yet the domain of Wikipedia is very wide and not all categories are geography-related. Therefore, UGSC tags (presented in Section 2.4) can be applied as filters to extract relevant Wikipedia categories.

The first step of our approach was to build a filter of geography-related terminology. For our example, we gathered existing UGSC tags from the most popular UGSC projects (OSM, Wikimapia, and GMM). These tags were organized following the GMM theme classification, i.e. natural features, roads, cities and political regions, and POIs. This scheme seems more intuitive than that of OSM. For extracting these tags, the main difficulty was that they can only be manually extracted from help pages. A set

of 432 tags were obtained; there are 66 tags for nature-related elements, 95 for roads and networks, 21 for administrative-related items, and 250 for POIs. This process is illustrated in Figure 1.

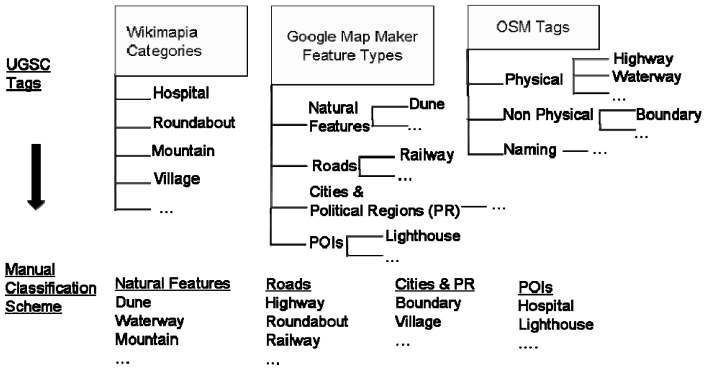

Fig. 1. Step 1: The classification process for an excerpt of UGSC tags (for this paper all tags were translated to the English language)

Importantly, these tags are used to run our process, both to provide initial specification elements and to illustrate the process, but any keyword provided by a user could be used instead of these tags.

The second step consisted in creating features types by querying Wikipedia using the filter described above – or user keywords in the future. We extracted categories and subcategories of these categories. For instance, the category road infrastructure contains the subcategories roads, road bridges, rail trails, road junctions, and pedestrian crossings. Then, for every category, we extracted the corresponding infobox if available. Besides extracting categories, we also retrieved Wikipedia articles for every UGSC tag. They may be considered feature types as well. Indeed, the distinction between categories and articles in Wikipedia should not be systematically interpreted as a distinction between classes and instances (Zirn et al. 2008). For example, a highway would be considered as a feature type, not a geographic feature. Yet, in Wikipedia, there is an article and not a category for highway. Therefore, feature types correspond either to an article or a category in Wikipedia. For our example, the numbers of feature types obtained from Wikipedia using UGSC tags as a filter are presented by theme in Table 1.

The extraction process was carried on by performing string comparison between UGSC tags and titles of Wikipedia pages (and all string comparison of our process), we chose the N-gram similarity measure (Euzenat and Shvaiko 2007), with N=3 considering that most UGSC tags and Wikipedia

pages' titles are usually of small length. Wikipedia data can be manipulated by parsing huge XML database dumps of the entire encyclopedia.

Table 1. Number of feature types created from UGSC tags

	C&PR	NF	Roads	POIs
# UGSC Tags	21	66	95	250
# Feature Types	20	57	37	166

```
IN: theme_j: sets of UGSC tags by theme
OUT: UGSCModel (FTs (ATs, nmaFT, WikiSupCat, WikiSubCat),
                RTs): a UGSC model
Initialize ugscModel ← {}
Initialize currentFT, matchedIB, currentNMAFT ← ''
Initialize currentATs, currentSupCat,
    currentSubCat, currentNMAATs ← {}
for all themes t_i in theme do
  for all tag tag_j in t_i do
    currentFT ← getWikiPage(tag_j)
    currentSupCat ← getWikiSuperCategories(currentFT)
    setWikiSupCat(currentFT, currentSupCat)
    currentSubCat ← getWikiSubCategories(currentFT)
    setWikiSubCat(currentFT, currentSubCat)
    matchedIB ← getWikiInfobox(currentFT)
    currentATs ← getWikiInfoboxAttrList(matchedIB)
    setATs(currentFT, currentATs)
    currentNMAFT ← getNMAFT(currentFT)
    setNMAFT(currentFT, currentNMAFT)
    currentNMAATs ← getNMAATs(currentNMAFT)
    setNMAATs(currentFT, currentNMAATs)
    add(currentFT, ugscModel)
  end for
  for all ft ft_k in ftlist do
    rts ← getHypernymyWordNet(ft_k) +
          getMeronymyWordNet(ft_k)
  end for
  add(rts, ugscModel)
end for
return ugscModel;
```

Fig. 2. Step 2-4: Simplified Algorithm of the process for building a UGSC Model

Considering that we only need information about pages' titles and links between pages, we only queried three relational tables, pages, category, and categorylinks available as SQL dump files (state of October 2010). Querying these tables instead of the XML file solves the difficulties of handling large volumes of content. Nonetheless, the three tables are large in volume as well. Therefore, for optimizing the access to these tables, we created a SQL script which executes delete statements to erase tuples contained in administrative namespaces (e.g. projects, users, etc.). We also tested several indexing structures by measuring processing time and number of disk-block access. These tests showed us that the indexing structures proposed

by Mediawiki provide a reasonable query processing time. This first step of building feature types is summarized in lines 10–15 of a simplified version of the algorithm for the proposed process (Figure 2).

The third step consisted in looking for attributes for our newly created feature types. Wikipedia infoboxes are an important source of attribute-level information. Most infoboxes are retrieved through Wikipedia categories. Yet, there are some infoboxes that are associated only to articles and not to categories. For the human settlement infobox there is both an article and an infobox, but not a category. Next, every attribute specified in the matched infoboxes is assigned to the corresponding feature type. There are two clear issues at this point. First, syntactically similar attributes are repeated in these infoboxes. In this case, a simple merge based on string comparison can help solve it. Second, there are some attributes syntactically different but semantically similar. For instance, state and region are both the primarily administrative division in Germany and France, respectively. Wu and Weld (2009) have built a refined infobox ontology for the English Wikipedia which solves some of these issues. In future work, we plan to include this ontology by automatically translating its concepts using WikiNet (Nastase et al. 2010), which is a multilingual concept network built from Wikipedia.

Infoboxes are incrusted in articles using Wikicode. For instance, the infobox for articles related to rivers is *{{Infobox river| name=val1| ... |river_system=valn}}*, where *val{1...n}* are optionally provided by contributors. Instead of extracting infoboxes' content from the XML Wikipedia dump file, we used the raw infobox data set provided by DBpedia, especially considering that it is the only available information about infoboxes provided in the French language. We retrieved 746 infoboxes from the DBpedia dump (state of March 2010). For our example, we were only able to automatically retrieve 53 relevant infoboxes, leaving more than half of the feature types with no attribute types. This step of building attribute types is summarized in lines 16–17 of the algorithm in Figure 2.

At this step, we also retrieved feature types and attribute types from the model of a specific NMA topographic large scale database. This information was available as a geographic ontology of topographic concepts (Abadie 2009). The detailed description of geometry and attribute types was also available in XML format. In this way, a feature type points to a reference feature type from a NMA model and also contains attribute types retrieved from this reference model. This step is summarized in lines 18–21 of the algorithm in Figure 2. Retrieving these items is interesting with respect to two functions. The first function is to see how an NMA structures a given category of features and to possibly check specific integrity con-

straints. The second function is that the community can use the NMA features as "external references instances" in its model. For instance, the user through the editing GUI could establish a relationship "is within" between a user-defined feature type "restaurant" and an NMA feature type "building." This relationship can be used during edition to check if the restaurants are actually located in buildings.

3.2 Relationship types for UGSC

The fourth step consisted of acquiring relationship types. For this, we have firstly explored the Wikipedia article and category structure. The Natural Language Processing (NLP) community has built two Wikipedia graphs (Zesch and Gurevych 2007): the Wikipedia Category Graph (WCG) and the Wikipedia Article Graph (WAG). The authors provide the following definition: Wikipedia articles form a network of semantically related terms and constitute a direct graph where each node is an article and an edge is an explicit link between articles. Wikipedia categories are organized in a taxonomy-like structure; each category can have an arbitrary number of subcategories where a subcategory is typically established because of a hyponymy or meronymy relationship. However, Hecht and Raubal (2008) explain that most of the relationships in the WCG are limited to hyponymy relations with a sprinkling of meronymy relations.

Therefore, we looked for a more appropriate source to acquire relationship types for clarifying the semantics of the relationship types and for precisely distinguishing composition and specialization relationships. This kind of relationship corresponds to a lexical relationship, i.e. meronymy and hypernymy, respectively. One of the most used resources for discovering lexical relationships between words is WordNet (Miller 1995)[5]. It is a freely available dataset developed for the English language. It has been widely exploited by the research community and integrated in popular dictionary-based websites as a linguistic support (e.g. The Free Dictionary). For the implementation, we used EuroWordNet[6], the European version of WordNet. It has been built by linguistic experts by automatically translating the English version (WordNet 1.5) in a European language like French. This version contains around 22500 synsets shared out between nouns and verbs, including a manual verification. Words or synsets are interlinked by means of conceptual-semantic and lexical relations, such as hypernymy/hyponymy, meronymy/holonymy, synonymy, and antonymy. By

[5] http://wordnet.princeton.edu
[6] http://www.illc.uva.nl/EuroWordNet

exploring this resource, we were able to extract relationship types among the entire set of feature types extracted in the previous step. We were particularly interested in hypernymy and meronymy relationships which play an important role when evaluating spatial integrity constraints on geographic data. This step of building relationship types is summarized in lines 24–28 of the algorithm in Figure 2. An excerpt of the meronymy relationships created for the proposed example is illustrated in graph form, in Figure 3. The resulting directed graph consists of nodes and edges representing feature types and relationship types (i.e. hypernymy/meronymy relationships), respectively. For instance, *sidewalk* → *road* means that roads are composed of sidewalks.

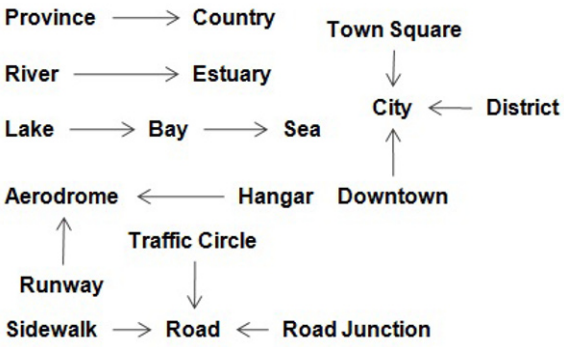

Fig. 3. Excerpt of meronymy relationships created for the proposed example (tags were translated to the English language)

3.3 Discussion

In general, we have found several issues of structure and consistency from Wikipedia, which are being investigated by the NLP community. Therefore, our approach inherits some of these limitations. Notably, this community needs to tackle particularities of languages different than English.

The proposed approach is also limited by WordNet coverage of relationships. This may be solved by adding another expert-validated source of information which can provide new relationship types to the UGSC model. A solution may be provided by Cyc[7], which is a large scale knowledge repository of everyday common sense knowledge. Concerning spatial content, an issue in WordNet is the relatively small amount of meronymy relationships with respect to hypernymy ones. That is, there is a large number

7 http://www.cyc.com

of specialization relationships and a small amount of composition relationships. It is unfortunate since the latter are of high importance for enhancing consistency of the content when evaluating spatial integrity constraints.

Another issue is the relatively small amount of Wikipedia infoboxes. That is, there is no guaranty that all matched categories or articles will have infoboxes. For improving infobox templates, the French Wikipedia has created the project Infobox Version 2[8]. They expect to enhance the definition of infoboxes, increase their coverage, and merge redundant infoboxes. This project can bring light to our issue of an insufficient number of attribute types. We have also considered to translate to the French language the refined infobox ontology provided by Wu and Weld (2009). This will allow solving the issue of insufficient coverage of Wikipedia infoboxes. The concepts of this ontology will be automatically translated using WikiNet (Nastase et al. 2010), which is a multilingual concept network built from Wikipedia.

4 Conclusion and future work

In this paper, we presented a novel proposition for managing the quality of UGSC based on enhanced specifications. To assist contributors in building such specifications, we have developed a generic process for structuring UGSC by yielding relevant modeling elements from user keywords. These elements are feature types, attributes types, and relationships types. For creating feature types, Wikipedia articles and categories are retrieved by applying a geography-related filter, which is derived from UGSC tags. Afterwards, for every newly created feature type, infoboxes, super- and subcategories are extracted from Wikipedia. Next, attribute types are created for every attribute of the matched infoboxes. For acquiring relationship types, the lexical relationships hypernymy and meronymy are queried in WordNet for every newly created feature types. Feature types, relationship types, and integrity constraints from an NMA data set are also retrieved, that can be used both to get suggestions about how to structure a particular content and to make relationships between UGSC and NMA content. Our proposed process allows users to take the best from two worlds when structuring their content- the world of UG(S)C and the world of conventional geographic databases. The preliminary model obtained from current tags of UGSC projects will be available on demand. These results can then be evaluated or compared to other UGSC proposed specifications. In fu-

[8] http://fr.wikipedia.org/wiki/Projet:Infobox/V2

ture work, we plan to perform user tests to investigate whether the proposed method actually helps users to build a model for their spatial content. Moreover, we will investigate the reconciliation of distributed operations on UGSC.

Acknowledgements

The authors are grateful to the reviewers of the submitted version of this paper for their most valuable inputs and comments to improve it.

References

Abadie N (2009) Formal Specifications to Automatically Identify Heterogeneities, in the 12th AGILE International Conference on Geographic Information Science Pre-Conference Workshop: Challenges in Spatial Data Harmonization, Hannover, Germany

Antoniou V, Haklay M, Morley J (2010) A step towards the improvement of spatial data quality of Web 2.0 geo- applications: the case of OpenStreetMap, in the 18th GISRUK Conference, London, UK, pp. 197–201

Bédard Y, Larrivée S, Proulx MJ, Nadeau M (2004) Modeling Geospatial Databases with Plug-Ins for Visual Languages: A Pragmatic Approach and the Impacts of 16 Years of Research and Experimentation on Perceptory. In: Wang S et al. (eds) Conceptual Modeling for Geographic Information Systems Workshop, LNCS 3289, pp. 17–30

Bishr M, Kuhn W (2007) Geospatial Information Bottom-Up: A Matter of Trust and Semantics, in: Fabrikant SI, Wachowicz M (eds) The European Information Society - Leading the Way with Geo-information, Springer Verlag LNGC, pp 365–387

Budhathoki N R, Nedovic-Budic Z, Bruce B (2010). A framework for volunteered geographic information: Proposal and illustration. Geomatica 64 (1): 11–26

Brando C, Bucher B (2010) Quality in User Generated Spatial Content: A Matter of Specifications, in the 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal

Bruns HT, Egenhofer MJ (1996) Similarity of Spatial Scenes, in Proceedings of the 7th International Symposium on Spatial Data Handling, pp 31–42

Euzenat J, Shvaiko P (2007) Ontology Matching, Springer-Verlag, Berlin Heidelberg, p. 73

Gilles J (2005) Internet encyclopedias go head to head. Nature 438(7070): 900–901

Girres JF, Touya G (2010) Quality Assessment of the French OpenStreetMap Dataset. Transactions in GIS 14 (4): 435–459

Goodchild M (2007) Citizens as Sensors: The World of Volunteered Geography. GeoJournal 69(4): 211–221

Graham M (2010) Neogeography and the Palimpsests of Place. TijdsCHrift voor Economische en Sociale Geograpfie 101(4): 422–436

Haklay M (2010) How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England. Environment and Planning B 37(4), pp. 682 – 703

Hecht B, Raubal M (2008) GeoSR: Geographically Explore Semantic Relations in World Knowledge, in Bernard L, Friis-Christensen A, Pundt H (eds) 11th AGILE International Conference on Geographic Information Science, Springer Verlag LNGC, pp 95–113

ISO (2003) Geographic Information - Metadata, International Standard, TC211/ ISO19115:2003

ISO (2005) Geographic Information – Rules for application schema, International Standard, TC211/ ISO19109:2005

Kavouras M, Kokla M (2008) Theories of geographic concepts – Ontological Approaches to Semantic Integration. CRC Press

Mäs S (2007) Checking the Integrity of Spatial Semantic Integrity Constraints, Constraint Databases, Geometric Elimination and Geographic Information Systems

Mihalcea R (2007) Using Wikipedia for Automatic Word Sense Disambiguation, in Proceedings of the North American Chapter of the Association for Computational Linguistics, Rochester, USA

Miller GA (1995) WordNet: A Lexical Database for English, Communications of ACM 38(11): 39–41

Nastase V, Strube M, Boerschinger B, Zirn C, Elghafari A (2010) WikiNet: A Very Large Scale Multi-Lingual Concept Network, in Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta

Oster G, Urso P, Molli P, Imine A (2006) Data consistency for P2P collaborative editing, in Proceedings of the ACM Conference on Computer-Supported Cooperative Work, pp. 259–267, Alberta, Canada

Turner A (2006) Introduction to Neogeography, O'Reilly Media, p 2

Parent C, Spaccapietra S, Zimanyi E, Donini P, Plazanet C, Vangenot C (1998) Modeling Spatial Data in the MADS Conceptual Model, in Proceedings of the 8th International Symposium on Spatial Data handling, Vancouver, Canada, pp. 138–150

Wu F, Weld DS (2008) Automatically Refining the Wikipedia Infobox Ontology, in Proceedings of the 17th International World Wide Web Conference, Beijing, China, pp. 635–644

Zesch T, Gurevych I (2007) Analysis of the Wikipedia Category Graph for NLP Applications, in Proceedings of the TextGraphs-2 Workshop, pp 1–8

Zirn C, Nastase V, Strube M (2008) Distinguishing between Instances and Classes in the Wikipedia Taxonomy, in Bechhofer S, Hauswirth, Hoffmann J, Koubarakis (eds) The Semantic Web: Research and Applications, 5th European Semantic Web Conference, Springer Verlag LNCS, pp 376–387