Enhancing the online discovery of geospatial data through taxonomy, folksonomy and semantic annotations

Samy Katumba, Serena Coetzee

Centre for Geoinformation Science (CGIS), Department of Geography, Geoinformatics and Meteorology, University of Pretoria, South Africa, skabangu@hotmail.com

DOI: http://dx.doi.org/10.4314/sajg.v4i3.14

Abstract

Spatial data infrastructures (SDIs) are meant to facilitate dissemination and consumption of spatial data, amongst others, through publication and discovery of spatial metadata in geoportals. However, geoportals are often known to geoinformation communities only and present technological limitations which make it difficult for general purpose web search engines to discover and index the data catalogued in (or registered with) a geoportal. The mismatch between standard spatial metadata content and the search terms that Web users employ when looking for spatial data, presents a further barrier to spatial data discovery. The need arises for creating and sharing spatial metadata that is discoverable by general purpose web search engines and users alike. Using folksonomies and semantic annotations appears as an option to eliminate the mismatch and to publish the metadata for discovery on the Web. Based on an analysis of search query terms employed when searching for spatial data on the Web, a taxonomy of search terms is constructed. The taxonomy constitutes the basis towards understanding how web resources in general, and HTML pages with standard spatial metadata in particular, can be documented so that they are discoverable by general purpose web search engines. We illustrate the use of the constructed taxonomy in semantic annotation of web resources, such as HTML pages with spatial metadata on the Web.

1. Introduction

Spatial data infrastructures (SDIs) are meant to facilitate dissemination and consumption of spatial data, amongst others, through publication and discovery of spatial metadata in geoportals. However, geoportals are often known to geoinformation communities only and present technological limitations which make it difficult for general purpose web search engines to discover and index the data catalogued in (or registered with) a geoportal. In addition, spatial metadata is created without knowledge of how the actual documented data will be searched and discovered by users. The mismatch between spatial metadata standard content and the search terms that web users employ when looking for spatial data online, presents a further barrier to spatial data discovery. The need arises for creating and sharing spatial metadata that is discoverable by general purpose web search engines and users alike.

To this end, a study was conducted to analyse search terms users employed when searching for spatial data. The study results informed the construction of a taxonomy comprising of categories (classifications) of the above search terms. Such a taxonomy is useful for consideration in the spatial metadata content creation by spatial data producers. Furthermore, to enhance the discovery of spatial metadata resources (HTML web pages) on the Web, this study proposes the semantic annotation of search terms with terms from an ontology for which a mapping to spatial metadata exists. The ontology describes geographic phenomena or features represented by the spatial data in question. The proposed solution overcomes the challenges hindering the discovery of spatial metadata (data) in main stream information retrieval.

The research reported in this paper is part of an experimental endeavour to empirically test the discoverability of HTML pages with information about spatial data resources by general purpose web search engines. It is motivated by current difficulties in finding spatial data with general purpose web search engines (Katumba and Coetzee 2013).

The remainder of the article is structured as follows: relevant concepts and related work are described in section 2; section 3 presents the methodology applied in this research as well as the experiment set up. A detailed discussion of the results is provided in section 4. Section 5 concludes the paper.

2. Background and related work

2.1 Folksonomy and Logsonomy

The word ‘folksonomy’ is derived from the combination of ‘folk’ and ‘taxonomy’. A folksonomy is constructed when users (participants) attach their own keywords (tags) to online resources, such as web pages and videos, as a means of describing such resources on online social (or collaborative) tagging platforms (Trant 2008). Since 2004, the number of social tagging platforms has been increasing. Popular examples are Flickr and Delicious. A folksonomy eases the search and discovery of online information (Lee and Young 2008).

A folksomy built from a corpus of keywords obtained from a web search engine log data is referred to as a 'logsonomy' (Jaschke *et al*. 2008). In this research a logsonomy is employed, given that the corpus of keywords used to build the taxonomy is obtained from two logs: 1) keywords used by participants searching for spatial data; and 2) keywords obtained from the Bing webmaster tools.

2.2 Semantic annotation

Letting users participate in the documentation of spatial data with their own keywords (tags) may be beneficial as users will be more likely to use the same terms during the retrieval process. In that perspective, Kalantari *et al*.(2010) proposes the use of folksonomies as a means of automatically enriching spatial metadata contents for improved discovery within SDIs.

However, the effectiveness of folksonomies can be hindered by ambiguous semantics or meanings in the keywords employed by users participating in the tagging process, for example, keywords are typically inconsistent and redundant because there are no rules guiding social tagging. As a result, researchers have suggested to associate semantics with tags (keywords) through the use of ontologies (Trant, 2008; Lee and Young, 2008). Such a process is known as semantic annotation. Semantic annotations establish a connection between the resource, its metadata, and an ontology (OGC 2012). The meaning of the terminology used in semantic annotations is commonly understood by being associated with a shared conceptualisation also known as 'ontology' (Oren *et al*. 2006).

In 2012, the Open Geospatial Consortium (OGC) published an OGC Best Practice document on semantic annotations in OGC standards. The document explains how providers should attach meaningful descriptions to OGC compliant data and services. It is recommended that ontologies are connected to OGC data and services at three levels namely: 1) the metadata level (in catalogues); 2) the data model and process description level; and 3) the level of the actual data instance in the database (OGC 2012). In this study, semantic annotation is applied by linking feature instances (resources) described in terms of ISO 19115:2003 metadata to concepts in an ontology. With this in mind, in this paper we propose a bottom-up approach (involving users) to annotating online spatial data resources. This complements the top-down approach described in the OGC Best Practice document, which only involves data custodians (creators).

2.3 Related work

Researchers have followed various approaches towards enhancing the discovery of web resources with spatial information, in particular through semantic annotation. Wu *et al.* (2006) applied statistical techniques on a collection of tags from 'delicious' (a social tagging online web application) to derive categories of semantics from these tags. This was done with the aim of constructing related ontologies from the tags employed by users in order to semantically annotate web resources. To find context information for tags used in Flickr (a social collaborative online platform for tagging pictures), Chen *et al*. (2008) used the Google web search engine's result contents to identify sentences (phrases) in which such tags appeared. Al-Khalifa *et al*. (2006) designed a tool (FolksAnnotation), which enables the mapping of folksonomy tags to concept terms from specific domains and resource type ontologies.

Sladic *et al*. (2011) suggest the linkage between an 'ontology based knowledge model in the field of real estate cadastre' and 'metadata and application schemas of the data resources' served by OGC web services. This came after they expressed concerns that OGC web services and standards lack semantic descriptions and are isolated from external knowledge models (ontologies). Lui *et al*. (2013) developed 'Geosearch', a search engine with the aim of improving the search of spatial metadata registered with the Global Observation System of Systems (GEOSS) clearinghouse. 'Geosearch' employs the Semantic Web for Earth and Environmental Terminology (SWEET) ontology.

A similar approach is adopted in the work described in this paper. However, what is different about this study, is that it starts with an analysis of search terms (tags) employed by users to specifically find spatial data. A taxonomy is constructed based on a logsonomy (as opposed to folksonomies used in other research). The taxonomy is a means of understanding the search terms (tags) and to facilitate the semantic annotation process. The proposed semantic annotation then draws on the taxonomy when linking search terms with corresponding terms in relevant ontologies.

3. Methodology

This paper presents a method for linking user search terms or folksonomy tags with associated concepts contained in shared conceptualised knowledge domains (ontologies). A logsonomy, a corpus of search terms obtained from experiments, is used to construct a taxonomy of search terms from which the proposed method for performing semantic annotation is developed. This is done based on the assumption that users are more likely to tag web resources on folksonomy platforms with the same terms they employ when searching for those resources using general purpose web search engines. Two experiments were performed:

1. User experiment

   Collection of search terms submitted to general purpose web search engines by human subjects (participants) who participated in a computer laboratory experiment in which they were instructed to search for spatial data.

2. Bing webmaster tools experiment

   Collection of a list of search terms that triggered the appearance of web pages with ISO 19115:2003 spatial metadata content prepared for the experiment. See Figure 1. The list was obtained from the log in the Bing webmaster tools (http://www.bing.com/toolbox/webmaster).

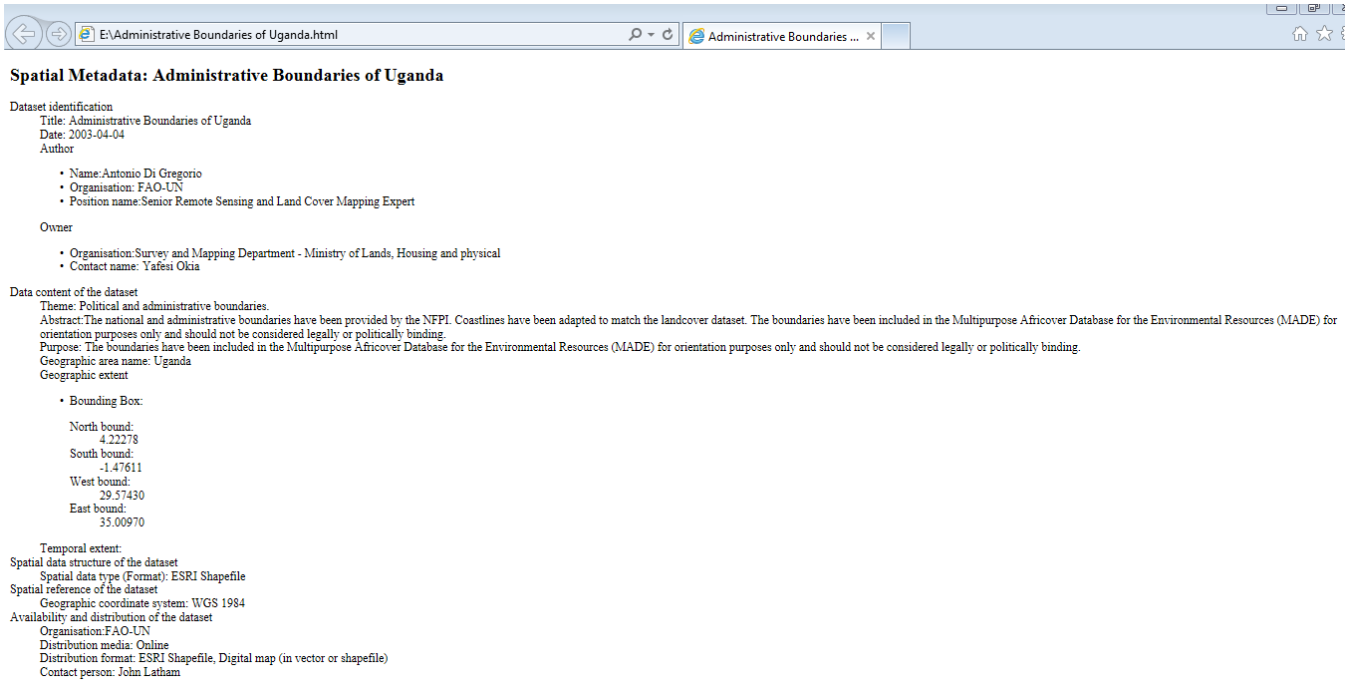

Figure 1. Example of a web page with ISO I9115:2003 spatial metadata content

After completion of experiments, an analysis of the collected search terms was performed in order to create categories of search terms for the taxonomy. The frequency of search terms was considered because it facilitates the creation of clusters of search terms to make up a category. The results of the analysis of the first experiment were compared to the results of the second experiment as a way of validating the taxonomy. The constructed taxonomy is comprised of categories (classes) of search terms, which can then be used for semantic annotation. The experiments were conducted in English only.

Since this experiment focussed on how users search for data and the metadata describing the data, the reliability and validity of the actual data was not evaluated.

Table 1 provides definitions for terminology used in this paper.

Table 1. Terminology

	Definition
Search term	Refers to a whole string of alphanumeric characters entered by a participant (web searcher), e.g. ‘namibia’ or ‘shapefile’.
Query	Is an occurrence of a single search term or a combination of more than one search term submitted to the web search engine. e.g.‘namibia shapefile’.
Session	Refers to the duration of a web search experience of a single user.

3.1. Experiment set up

3.1.1 First experiment: User experiment

Three groups of users participated in the experiment: 1) a group of 21 BSc Geoinformatics first year students from the University of Pretoria (UP); 2) a group of 17 BSc Geoinformatics students in their final year (third year) of studies, also from UP; and 3) a group of nine South African professionals in the field of geoinformatics. These three groups represent three different levels of expertise in geoinformatics. This choice was made with respect to the studies by Holscher and Strube (2000) who concluded that user expertise in the knowledge domain of the search influences success in web searches. Furthermore, a user-centered methodology for constructing the taxonomy is justified by the fact that users are the ones driving online searches when they search for spatial data. The experiment was repeated with each group on a different day and lasted for two hours. Participants were instructed as follows: "Search for spatial data to address a particular need (or to solve a given problem) of your choice for specific countries or cities in Africa. One should be able to visualize (and/or analyse) such spatial data on a GIS desktop software (in e.g. ArcGIS or QGIS)." They were advised to use any knowledge they had acquired throughout their studies and/or profession. The keywords used in their queries were monitored using Mozilla Firefox web browser cache.

3.1.2 Second experiment: Bing webmaster tools experiment

A set of 58 HTML web pages with information about spatial data (spatial metadata) was uploaded onto a web server (website) on a public domain (internet). An index page with a list of the 58 pages was also uploaded. The spatial metadata information contained in the HTML web pages was structured according to ISO 19115:2003. The website containing the 59 HTML web pages was registered with the webmaster tools of the Bing web search engine. The Bing webmaster tools includes reports on page traffic, search engine optimisation (SEO), crawl information and search keywords. The latter report provided a list of search terms that triggered the appearance of the uploaded HTML pages. A total of 90 queries was recorded in the Bing webmaster tools in the period starting on the 23rd November 2014 and ending on 18th February 2015, totalling 87 days.

Bing was chosen because it is a popular web search engine and makes available a comprehensive list of keywords that trigger the appearance of HTML web pages that are registered with the Bing web master tools. Alternatives, such as, Google webmaster tools, do not provide such a comprehensive list of keywords.

4. Results

4.1 Categories of the taxonomy

The construction of the taxonomy was informed by the hierarchy of different levels of granularity of metadata described in ISO 19115:2003 (Figure 2):

- Dataset series: a collection of spatial data with similar characteristics of theme, source date, resolution, and methodology, *e.g.* a collection of raster map data captured from a common series of paper, or a collection of vector datasets depicting surface hydrography with associated attribution for multiple administrative areas within a country.

- Dataset: a consistent instance of a spatial data product which may be composed of a set of identified feature types and instances, and attributes types and instances.

- Feature type: are feature constructs with common characteristics, *e.g.* all bridges within a dataset.

- Feature type instance: features (spatial constructs) that have a direct correspondence with a real world object, *e.g.* The Nelson Mandela bridge in Johannesburg.

- Attribute type: a digital parameter that describes a common aspect of grouped spatial primitives such as 0-dimensional (points), 1-dimensional (lines), 2-dimensional (polygons) and 3-dimensional geometric objects, *e.g.* overhead clearance associated with a bridge.

- Attribute instance: a digital parameter that describes an aspect of a feature instance, *e.g*. the overhead associated with a specific bridge across a road.

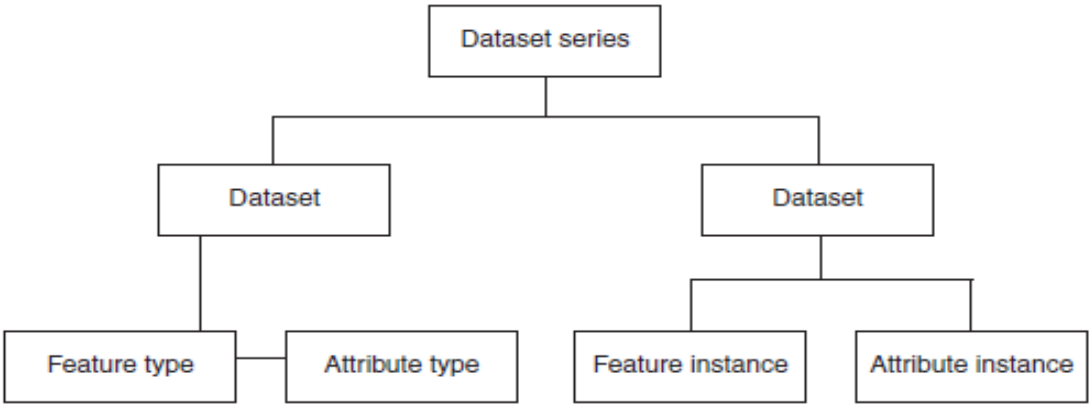

Figure 2. Hierarchy of metadata (Source: ISO 19115:2003)

Each query was broken into individual terms and each term was placed within a suitable category. The results suggest that web search terms employed by participants when searching for geospatial data using general purpose web search engines can be classified into four main categories (or clusters), which are defined based on selected core elements of ISO 19115:2003 spatial metadata standard as follows:

- ***Application domain***: This category defines the subject, the theme of the general topic of interest for which the geospatial data is being searched. Query terms (keywords) that fall under this category could be related, for example, to infrastructures, urban planning, water resources topics or subjects for which a given user is searching. ISO 19115:2003 geographic metadata elements associated with this category are 'Dataset topic category', 'Dataset title', 'Abstract describing the dataset' and 'Lineage'. Search terms falling under the Application Domain category fit one of these four metadata elements.

- ***Location***: This category defines the spatial extent or geographic coverage for which the spatial data in question is being searched. Query terms (keywords) related to location are place names or location types or classes. Examples are 'Africa', 'Johannesburg', 'Paris', 'Sub-Sahara', 'South America', etc. as place names and 'city', 'country', etc. as location types. This definition also includes land demarcations, such as municipalities or district. ISO 19115:2003 geographic metadata elements associated with this category are 'Geographic location of the dataset'. The standard provides for the location to be described by four (geographic) coordinates or by a geographic identifier. None of the search queries from the experiments contained geographic coordinates. Search terms falling under the Location category fit, the 'EX_GeographicDescription' metadata element.

- ***Feature type***: This category defines the actual geographic object, feature or phenomenon of interest to the user performing the search. Examples of search terms that fall under this category include roads, buildings, bridges and power lines. The ISO 19115 metadata element associated with this category is 'Spatial representation type'.

- ***Data model***: This category relates to the data representational model that spatial data employ. Spatial data can be represented as vector or raster. Under the vector model, query terms (keywords) display characteristics of primitive geographic features such as point, line or polygon. A common search term that falls under this category is 'shapefile', referring to the well-known vector (spatial) data representation model. The term 'raster' is used when searching for continuous geographic features such as rainfall, temperature and forests. The ISO 19115 metadata element associated with this category is 'Distribution format'.

As an illustration, Table 2 shows how some sample queries that were dissected and categorised. For example, 'Ghana land cover shape file' has five search terms, two fall under the 'Application Domain' category, one under the 'Location' category, none under the 'Feature type' category and two under the 'Model' category. The 'Ignored' column shows the number search terms that were not classified under any of the four categories. Examples for such terms include: adjectives, pronouns, conjunctions, adverbs and verbs, as well as words such as 'free', 'download', 'spatial', 'data', etc.

Table 2. Search term categorisation

Search terms	Total	Ignored	Category			
			Application Domain	Location	Feature Type	Data Model
'Ghana land cover shape file'	5		2	1		2
'uganda administrative boundaries shapefile'	4		1	1	1	1
'tanzania infrastructures shapefiles'	3		1	1		1
'south african roads and shapefiles'	5	1		2	1	1

Note: Search terms are included as entered by users, i.e. with proper nouns in lower case.

4.1 Results of the user experiment

Table 3 provides the analysis results for each of the groups of participants in the user experiment as well as the overall result. These results are discussed in subsequent sections.

Table 3. User experiment

Group	Total number of search terms	Ignored	Category			
			Application Domain	Location	Feature Type	Data Model
First years	866	510 (59%)	76 (9%)	150 (17%)	59 (7%)	71 (8%)
Third years	363	107 (29%)	76 (21%)	90 (25%)	35 (10%)	55 (15%)
Professionals	201	93 (46%)	44 (22%)	37 (18%)	18 (9%)	9 (5%)
Overall	1430	710 (50%)	196 (14%)	277 (19%)	112 (8%)	135 (9%)

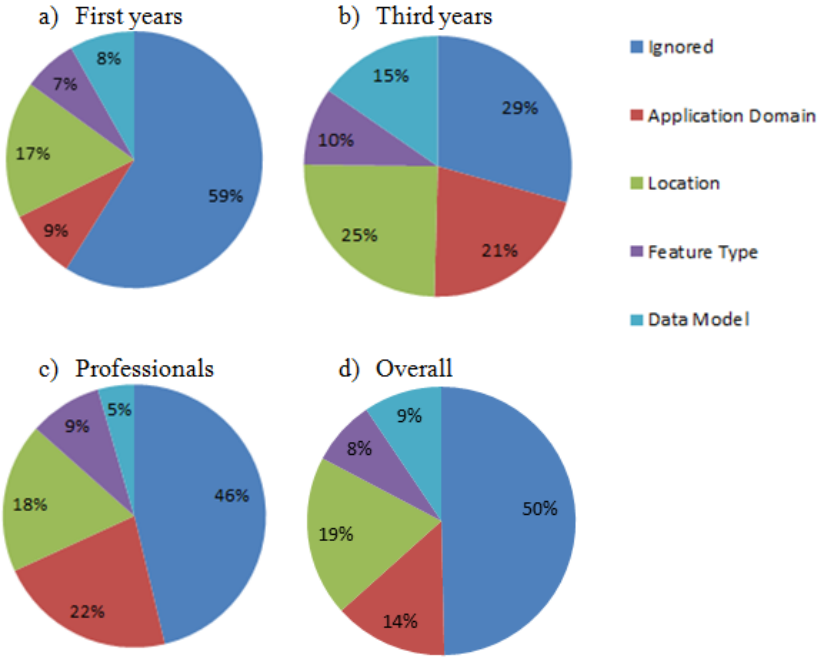

Figure 3. User experiment results

4.1.1 First year geoinformatics students

Being the largest group (21 participants) in the experiment, the first years submitted a total number of 159 unique search queries that were dissected into 866 search terms. Table 3 and Figure 3a show the results.

4.1.2 Third year geoinformatics students

A total number of 363 search terms were obtained from 88 unique search queries submitted to web search engines by 15 third year Geoinformatics students. As described in Table 3 and Figure 3b, 21% of the terms fell under the 'Application Domain' category, 25% of the terms fell under the 'Location' category, 15% of the terms fell under the 'Data Model' and 10% of the 'Feature Type'. 29% of the terms that fell under the 'Ignored' category were not related to any of the four categories of the taxonomy.

4.1.3 Professionals

A total number of 47 unique search queries were obtained from a group of nine professionals specialising in the field of geoinformatics (GIS). A total number of 201 search terms were derived from these search queries as categorised in the Table 3. As it can be seen from Figure 3c, 22% of the query terms employed by professionals were related to the application domain or subject for which the search for spatial data was being quested. 18%, 9% and 5% of the query terms were related to 'Location', 'Feature Type' and 'Data Model' respectively. 46% of the query terms that fell in the 'Ignored' were not related to any of the category described in the taxonomy.

4.1.4 Summary of user experiment results

A consolidated set of results of the search terms collected from the three different groups of participants (first year geoinformatics students, third year geoinformatics students and geoinformatics professionals) is summarised in Table 3. The 1430 search terms were derived from a total number of 294 unique search queries participants entered in the text search boxes of general purpose web search engines of their choice. From Figure 3d, it can be seen that 50% of the terms employed by participants fall under the 'Ignored' category leaving the other 50% shared between the four categories of the taxonomy. The category that had a high frequency of terms (19%) is the 'Location', followed by the 'Application Domain' category with 14% of terms. The 'Feature Type' category and the 'Data Model' category had 8% and 9% of the terms respectively.

The ratios of the number of participants to the number of search queries employed by the three groups are provided in Figure 4. The group of professionals has the highest ratio (0.19), followed by the third years (0.17) and the first years with the smallest ratio (0.14). These ratios can be expressed in terms of ten participants: ten first year students submitted 14 queries, ten third year students submitted 17 queries, and ten professionals submitted 19 queries. A possible explanation could be that the first years had to submit a higher number of queries to web search engines due to the fact that they were not obtaining desirable search results. Such a fact could be linked to the observation by Holsche and Strube (2000) that user expertise in the knowledge domain of the search influences success in web searches.

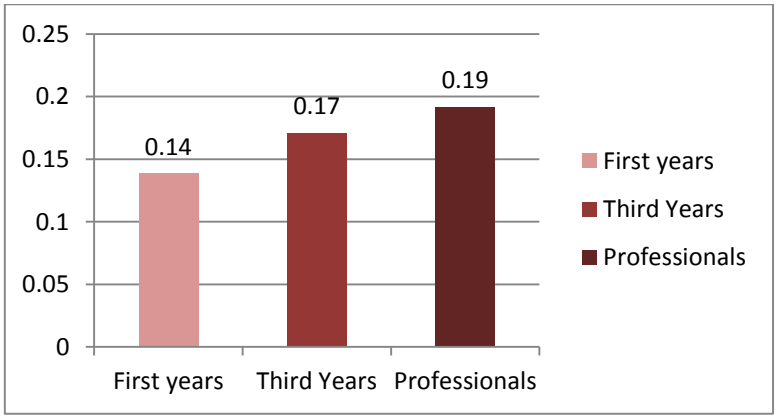

Figure 4. Ratio of number of participants to number of queries

Another comparison of ratios of the number of participants to the number of search terms as depicted in Figure 5 and Figure 6 (with and without the terms that fall in the 'Ignored' category) was performed. In Figure 8 the group of professionals has the highest ratio (0.045), followed by the third years group (0.041) and the first years group with the smallest ratio (0.025). These ratios can be expressed in terms of ten participants: ten first year students used 2.5 search terms, ten third year students used 4 search terms, and ten professionals used 4.5 search terms.

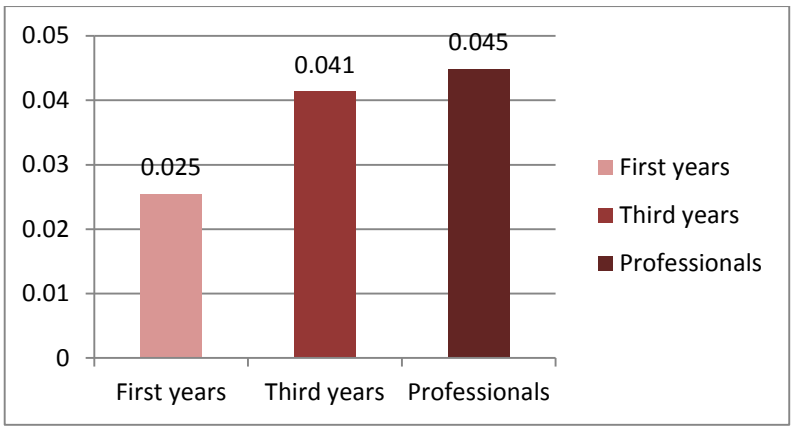

Figure 5. Ratio of number of participants to number of search terms (including terms in 'Ignored' category)

However, when looking at the results depicted in Figure 6, it can be seen that the third year group has the lowest ratio as compared to the two other groups. These ratios can be expressed in terms of ten participants: ten first year students used 6 search terms, ten third year students used 6 (5.9) search terms, and ten professionals used 8 search terms. One can only speculate why the third year students used less terms that could to be ignored. Research suggests that searchers by users who are knowledgeable in a field are more successful (Holsche and Strube 2000).

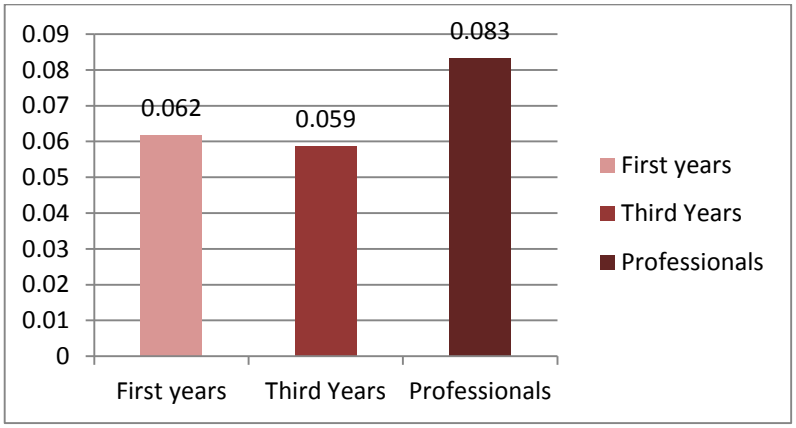

Figure 6. Ratio of number of participants to number of search terms (excluding terms in 'Ignored' category)

4.2 Results of the Bing webmaster tools experiment

A total number of 90 search queries were obtained from the Bing web search engine webmaster tool. After dissecting each of these queries as depicted in Figure 7, 23% of the terms fell under the 'Location' category, 18% of the terms fell under the 'Application Domain' category, 9% of the terms fell under the 'Feature Type' category and 5% of the terms fell under the 'Data Model' category. The 'Ignored' category was made of 45% of the total number of terms.

Table 4. Experiment results from the Bing webmaster tools

Total number of search terms	Ignored	Category			
		Application Domain	Location	Feature Type	Data Model
330	148 (45%)	58 (18%)	77 (23%)	30 (9%)	17 (5%)

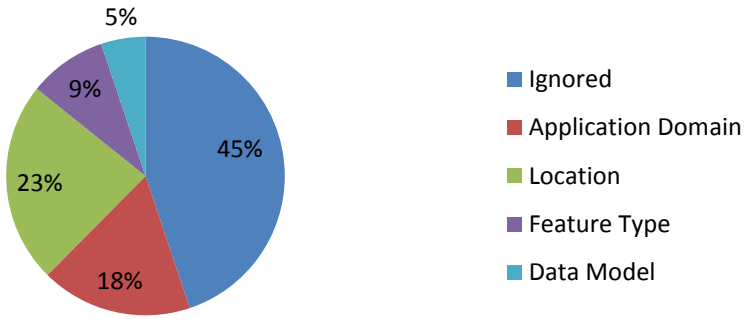

Figure 7. Experiment results – Bing webmaster tools application

4.3 Proposed taxonomy

It can be seen from the pie diagrams of both sets of results (Figure 3d and Figure 7) that the 'Location' category has the highest proportion of terms, followed by the 'Application Domain' category. This reveals the strong association with location in search queries with the intent of finding spatial data. The 'Application Domain' with the second highest number of search terms, demonstrates that the quest for spatial data is typically guided by the need to use such data in a given application domain. The 'Feature Type' and the 'Data Model' categories had the lowest proportions of terms in both sets of results as compared to the other two categories.

Summaries of both sets of results as illustrated in Table 5 reveal that the average number of terms per unique query of participants in the user experiment is 5 and the average number of terms per unique search query from the Bing webmaster tools is 4. In both sets of results, the average number of terms related to each of the 4 categories ('Application Domain', 'Location', 'Feature Type' and 'Data Model') is 1.

Table 5. Average number of terms per category

	Number of terms per query	Category				
		Ignored	Application Domain	Location	Feature Type	Data Model
User Experiment	5	3	1	1	1	1
Bing webmaster tools	4	2	1	1	1	1

The results of the two experiments lead one to conclude that on average, a search query employed with the intent of finding spatial data using web search engines, will be made of at least 1 term from the four categories provided in the taxonomy. Furthermore, on average, the longest search query is between 4 terms (user experiment conducted in the lab) and 5 terms (Bing webmaster tool). Such averages of the number of terms per query could be affected by a few outliers in each search query set. For example, while on average the number of terms per search query from the user experiment was 4, the set of search queries collected contained a query with 13 terms. Similarly, the longest search query from the Bing webmaster tools experiment contained 11 terms. Hence, the mode which provides the most frequently occurring value in a data set can be used as a descriptive measure for comparing the results of the two experiments. For example, the mode for both the individual experiments and the two experiments combined is three. This implies that a search query employed for finding spatial data on the web, is frequently comprised of three terms. Furthermore, given that the 'Application Domain' and 'Location' categories have the highest number or proportions of terms; search terms employed by users to find spatial data are more likely to fall under these two categories.

Based on the analysis of the search terms, a taxonomy for use in semantic annotation is proposed in Figure 8. The four categories of the taxonomy are shown in the rectangles, while ontologies (for the purpose of semantic annotation) are in ovals.

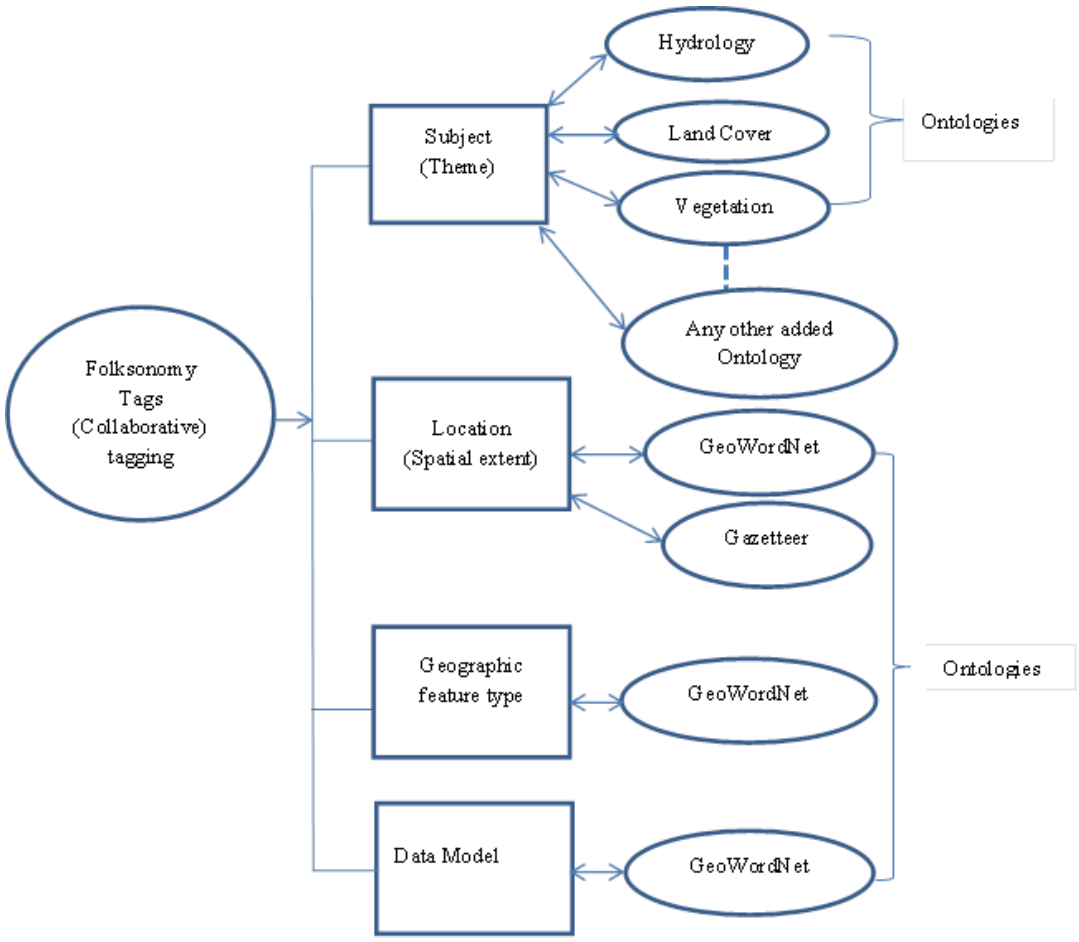

Figure 8. Taxonomy and semantic annotation

Note: GeoWordNet is a semantic resource (ontology) comprised of GeoNames (a 'geographical database that covers all country names and contains over eight million place names'),WordNet ('a large lexical database of English with nouns, verbs, adjectives and adverbs grouped into sets of cognitive synonyms, each expressing a distinct concept'), and other high quality resources (Giunchiglia *et al*. 2010).

4.4 Illustration of semantic annotation

Conceptualising the semantic annotation process illustrated in Figure 8 may appear as complex as its implementation. However, quite some work has already been done in terms of the design of geospatial standards to make the semantic annotations of geospatial web resources possible, for example, the OGC Best Practice document on semantic annotation of geospatial web resources by data custodians. In this paper we illustrated how users can also be part of this process by taking advantage of folksonomy tags to create semantic metadata.

In Figure 9, we demonstrate how the application profile for Dublin Core (DC) in social tagging proposed by Catarino *et al*. (2008) can be used in semantic annotation of geospatial resources. Catarino *et al*. (2008) identified 14 out 16 DC elements to which folksonomy tags could be assigned. However, the context in which the tags are used is not always obvious to assume, making it difficult to link folksonomy tags to DC elements. DC is used as reference description for online web resources (Dublin Core 2012) in main stream information retrieval where general purpose web search engines are widely used. The study described in this paper showed how the context of tags can be determined from an analysis of search queries based on core metadata elements defined in ISO 19115:2003. The existing mapping (CWA 14857:2003) between DC elements and ISO 19115:2003 provides the link to the DC elements.

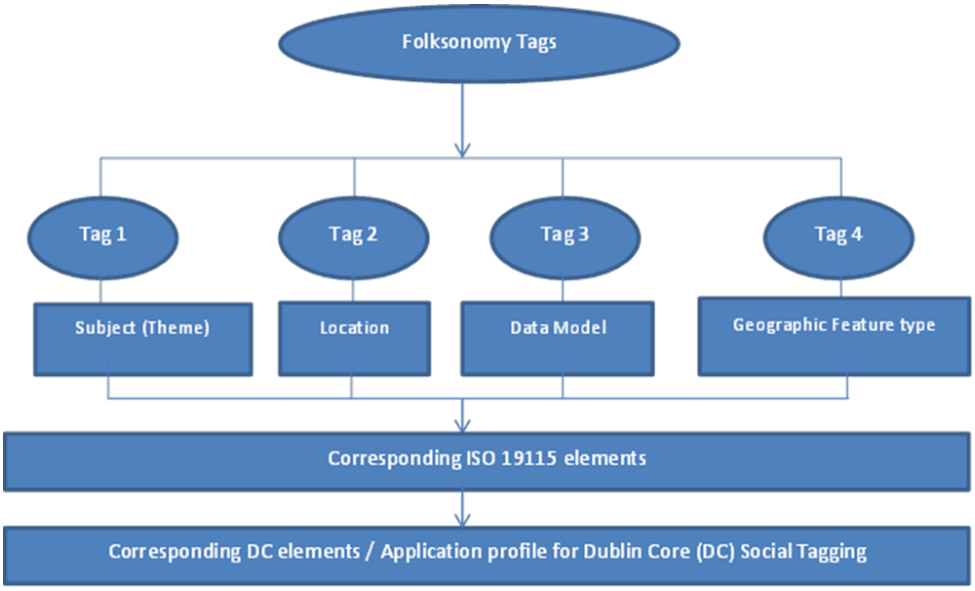

Figure 9. Method implementation

As illustrated in Figure 9, folksonomy (or logsonomy) tags are grouped into four categories: - the subject or theme of the search; the geographic location or the spatial extent in which the requested spatial data cover(s); the data model (raster or vector); and the geographic feature(s) that describe(s) the spatial representation of the real world feature being searched. The approach described in this paper proposes a way of enhancing the discovery of web resources (about spatial data) through main stream information retrieval by web search engines. By mapping user tags to expert terminology in ontologies, semantic annotation is led by the spatial consumers.

5. Conclusion

An analysis of a logsonomy of search terms employed when searching for spatial data using general purpose web search engines has been performed in the research described in this paper. The logsonomy was obtained from two experiments: one involving the capturing of search query terms employed by participants when searching for spatial data using web search engines, and the other involving the search terms in Bing webmaster tools that triggered the appearance of HTML pages with spatial metadata content.

The two experiments resulted in a logsonomy that informed the construction of a taxonomy of search terms for understanding the kind of search terms users employ when searching for spatial data on the web. The four categories in the taxonomy for which the collected search terms were grouped are: 'Subject (theme)', 'Location (spatial extent)', 'Geographic feature type' and 'Data Model'. The taxonomy thus constructed, serves in the illustration of the semantic annotation process whereby search terms (tags) employed by users are mapped to terms in ontologies for the purpose of associating them to their semantics or common meaning.

Besides providing insights to data producers in terms of compiling spatial metadata, this study is an example of involving the consumers in the process of describing spatial data published online. The results contribute towards enhancing the discovery of spatial data in main stream information retrieval where general purpose web search engines are commonly used.

One of the challenges of manually creating a taxonomy of search terms is the inconsistencies that may arise in the assignment of terms to appropriate categories. Future work should explore automatic categorisation assisted by human supervision. To verify these results from this exploratory study, the analysis of search terms performed through the use of descriptive statistics in this study could be supplemented by advanced inferential statistical analysis (such as ANOVA, etc.) of results from bigger search query sets, which may provide rigorous or other means of analysis and comparison of results. It would also be interesting to repeat the experiment in different languages to see whether there are any differences in the results.

6. References

Al-Khalifa, S & Davis C 2006, 'Folksannotation: A semantic metadata tool for annotating learning resources using folksonomies and domain ontologies'. *Innovations in Information Technology,* November 2006.

Al-Khalifa, S & Davis, C 2007, 'Exploring the Value of Folksonomies for Creating Semantic Metadata'. *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 3, no. 1, pp. 13-39.

Catarino, M. & Baptista, A 2008, 'Relating Folksonomies with Dublin Core', *In Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, Singapore.

Chen, M, Liu X & Qin J 2008, 'Semantic Extraction from Socially-Generated Tags: A Methodology for Metadata Generation', *In Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, Singapore.

Dublin Core Metadata Initiative (2012), http://dublincore.org/, viewed 7 May 2013

Gahegan, M & Adams, B 2013, 'The Many Semantic Domains of Spatial Data Infrastructure', *Workshop on Semantics in Geospatial Architectures: Applications and Implementation*, Pyle Center, 28-29 October 2013, University of Wisconsin-Madison.

Giunchiglia, F, Maltese, V, Farazi F, & Dutta B 2010, 'GeoWordNet: a resource for geo-spatial applications' *In Proceedings of the 7th international conference on The Semantic Web: research and Applications*, vol. Part I, pp. 121-136, Springer-Verlag Berlin, Heidelberg, 2010

Holscher, C & Strube, G 2000, 'Web search behavior of Internet experts and newbies', *Computer Network*, vol. 33, pp. 337-346.

Intagorn, S & Plangprasopchok A 2010, 'Harvesting Geospatial Knowledge from Social Metadata', *Proceedings of the 7th International ISCRAM Conferenc*e, Seattle, May 2010, USA, 2010.

Kalantari, M, Olfat, H, & Rajabifard, A 2010, ' Automatic metadata enrichment: reducing spatial metadata creation burden through spatial folksonomies', in A Rajabifard et al. (eds.), *Spatially enabling society,* pp. 119–129, Luven University Press, Luven, 2010.

Katumba S and Coetzee S 2013, 'Spatial data discovery using general purpose web search engines', *26th International Cartographic Conference*, Dresden, 25 - 30 August 2013, Germany.

Lee, S & Yong H 2008, 'OntoSonomy: Ontology-based Extension of Folksonomy', *IEEE International Workshop on Semantic Computing and Applications*, Seoul, Korea, 2008

Lui, X, Song, Y, Liu, S & Wang, H 2012, ' Automatic taxonomy construction from keywords', *KDD'12*, Beijing, August 12-16, China, 2012.

Lui, K, Yang, C. & Gui, Z 2013, 'GeoSearch: A system Utilizing Ontology and Knowledge Reasoning to Support Geospatial Data Discovery', *Workshop on Semantics in Geospatial Architectures: Applications and Implementation*, Pyle Center, 28-29 October 2013, University of Wisconsin-Madison.

Maué, P 2012, *Semantic annotation in OGC standards*, viewed 7 February 2013, http://www.opengeospatial.org/node/1790

Oren, E, Moller K, Scerri, S, Handschuh, S & Sintek, M 2006, *What are Semantic Annotations?, Technical report*, DERI Galway, viewed 7 February 2013, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.7985&rep=rep1&type=pdf

Sladic, D, Govedarica, M & Ristic, A. 2011, 'Semantic Metadata in Spatial Information Systems, *IEEE 9th International Symposium on Intelligent Systems and Informatics (SISY 2011)'*, September 2011, Subotica, Serbia

Sujatha, R & Roa B 2011, 'Taxonomy construction techniques – issues and challenges', *Indian Journal of Computer Science and Engineering (IJCSE)*, vol.2, no 5.

Trant, J 2008, 'Studying social tagging and folksonomy: A review and framework', *Special Issue on Digital and User-Generated Content*.

Wu, X, Zhang, L & Yu, Y 2006, 'Exploring Social Annotations for the Semantic Web', *Proceedings of the 15th international conference on World Wide (WWW 06)*, Edinburgh , May 23-26, 2006, Scotland