The quality of thesauri used in geographic metadata

Dayany Díaz-Corona
Universidad de Zaragoza
María de Luna, 1
Zaragoza, Spain
dayanydc@unizar.es

Javier Lacasta
Universidad de Zaragoza
María de Luna, 1
Zaragoza, Spain
jlacasta@unizar.es

Javier Nogueras-Iso
I3A, Universidad de Zaragoza
María de Luna, 1
Zaragoza, Spain
jnog@unizar.es

Abstract

Although the use of controlled vocabularies such as thesauri are highly recommended in geographic metadata to facilitate the discovery in Spatial Data Infrastructures, this use is not as common as expected. One of the reasons for not taking advantage of these thesauri may be that there are some issues in the quality of these resources. This work analyses the quality of some of the most frequently used thesauri in geographic metadata, recommending some possible improvements.

Keywords: geographic metadata, thesaurus, quality analysis, Spatial Data Infrastructures.

1 Introduction

In order to facilitate the discovery and monitoring of spatial resources, the different Spatial Data Infrastructure (SDI) initiatives arisen during the last years have encouraged the use of controlled vocabularies such as code lists, taxonomies or more formalized thesauri through its recommendations for metadata creation (Fugazza and Luraschi, 2012).

For instance, in the case of SDIs in Europe, the European INSPIRE directive (European Union, 2007) for promoting and improving the sharing of spatial data in Europe proposes the use of several controlled vocabularies through the different annexes of its metadata regulation (European Commission, 2008). In particular, in the case of the *Keyword* metadata element, the INSPIRE metadata regulation forces the use of one keyword, at least, to describe (through a geographic metadata record) the spatial data theme referred by a dataset that is created and published in response to the implementation of the directive in one of the member states of the European Union. Indeed, in order to implement this measure the European Environment Agency (one of the bodies in charge of coordinating the implementation of INSPIRE) forced an extension of the General Environmental Multilingual Thesaurus - GEMET (European Environment Agency, 2017) to include the names of these INSPIRE spatial themes as new concepts in the thesaurus. Additionally, the INSPIRE technical guidelines for the metadata implementing rules (Joint Research Centre, 2013), establishing the mapping between the metadata regulation and ISO 19115 metadata standard, recommend a minimum of two keywords in addition to the mandatory keyword, and if possible, selected from controlled vocabularies (see Requirement 16 and Recommendations 11 and 12 in the Implementing Rules Technical Guidelines).

Despite these recommendations, the current holdings of metadata records make little use of these vocabularies. For instance, we analysed the use of thesauri in the metadata catalogue of the Spanish Spatial Data Infrastructure (IDEE), containing 3,640 records in September 2016 describing spatial datasets or spatial data series (see table 1). Although a reference to GEMET – INSPIRE Spatial Themes is present in the majority of records (85%) because it is mandatory, only 56% of records contain a reference to a thesaurus concept different from the INSPIRE Spatial themes.

Table 1: Details of analysed thesauri

Thesaurus	Concepts	BT-NT	RT	Use in IDEE
INSPIRE Spatial Themes	34	0	0	85.11%
GEMET (v4.0)	5244	5332	1043	36.13%
EUROVOC (v4.1)	6649	6628	3542	16.02%
AGROVOC (v1.3)	32060	32035	962	23.30%

Probably, one of the reasons for not making an extensive use of thesauri is the limited availability of qualified human resources for manual cataloguing. But apart from that, another possible reason to prevent cataloguers from the right use of thesauri could be the existence of specific issues in its quality (Albertoni et al, 2016). The purpose of this paper is to analyse the quality of the main thesauri employed in geographic metadata according to the automatic method proposed in (Lacasta et al, 2016), which reports the syntactic and semantic quality of a thesaurus with respect to ISO 25964 standard (International Organization for Standardization, 2011).

The rest of the paper is structured as follows. Section 2 summarizes the features of the method applied for reporting the thesaurus quality. Section 3 shows the results of the experiments done with thesauri used in geographic metadata. The paper ends with some conclusions and outlook on future work.

2 Thesaurus quality analysis tool

According to ISO 8402 (International Organization for Standardization, 1994), the "quality" is a measure of excellence or a state of being free from defects, deficiencies and significant variations. A thesaurus is defined as a controlled set of terms used in an application domain and the relations between those terms such as synonymy, broader-narrower (BT-NT) or related terms (RT) relations. The international standard ISO 25964 for multilingual thesauri defines the rules that should be followed to create correct thesaurus with respect to mandatory and optional properties (preferred labels, definitions), structure of the content (charset, use of acronyms,…), rules to assure uniformity along the thesaurus, and proper use of properties and relations.

For the analysis of thesauri studied within the context of this work, we have used the automatic tool proposed by (Lacasta et al, 2016), which is able to process thesauri represented in SKOS (Simple Knowledge Organization System: a W3C initiative to provide an RDF-based representation of terminological ontologies) to detect the fulfilment of the rules established in ISO 25964. There are other tools like qSKOS, Skosify or PoolParty that also analyse quality issues of SKOS-based vocabularies (Suominen and Mader, 2014). However, these last tools are focused on structural and lexical checks and we wanted to check also the content, context and semantics.

The selected tool generates a report with a set of measures informing about the percentage of properties and relations evaluated as correct in different aspects. With respect to the properties of thesaurus concepts, the tool analyses their completeness, content, context and complexity as follows:

- The completeness analysis measures the degree in which mandatory properties like preferred labels or definitions are provided for each concept.
- The content analysis detects isolated anomalies in label texts. Using regular expressions, the tool identifies invalid values inside labels such as non-alphabetic characters, adverbs, initial articles or acronyms in preferred labels.
- The context analysis detects anomalies involving several labels such as the detection of duplicate labels or inconsistencies in the use of upper case and plurals. The tool makes profit of lexical analysis algorithms such as stemmers to detect some of these inconsistencies (e.g., plurals).
- The complexity analysis detects syntactically complex labels. Thanks to the use of Part Of Speech (POS) taggers, the tool detects the presence of prepositions, conjunctions and adverbs that increase unnecessarily the complexity of label texts.

As regards relations, the tool analyses the structure and semantics of relations as follows:

- The structural analysis requires identifying the completeness of BT/NT (broader term/narrower term) relations, the existence of BT/NT cycles, and the informativeness of Related Term (RT) relations. Whereas the BT/NT completeness analysis verifies that there are not orphan concepts, graph-traversing techniques are used to check that there are not cycles in BT/NT hierarchies. Last, the informativeness of RT relations is assured if they do not involve two concepts already connected through a BT/NT hierarchical path.
- The semantic validity of BT/NT relations is done through the alignment of the thesaurus with WordNet (Fellbaum, 2008) and DOLCE (Gangemi et al, 2003) ontologies in successive steps. On the one hand, Wordnet (and its multilingual version known as Open Multilingual Wordnet) is a lexical database that groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets). On the other hand, DOLCE is an upper-level ontology consisting of very abstract categories of concepts and relations, which can be used to analyze relations independently of the thesaurus domain. After the alignment, if the original thesaurus relations are compatible with the ones provided in one of these two ontologies, BT/NT relations are considered as correct.

The semantic validity of BT/NT relations is one of the main contributions of the tool. For a better understanding of this part,

Figure 1: Semantic correctness of BT/NT relations through WordNet, modified after (Lacasta et al, 2016)

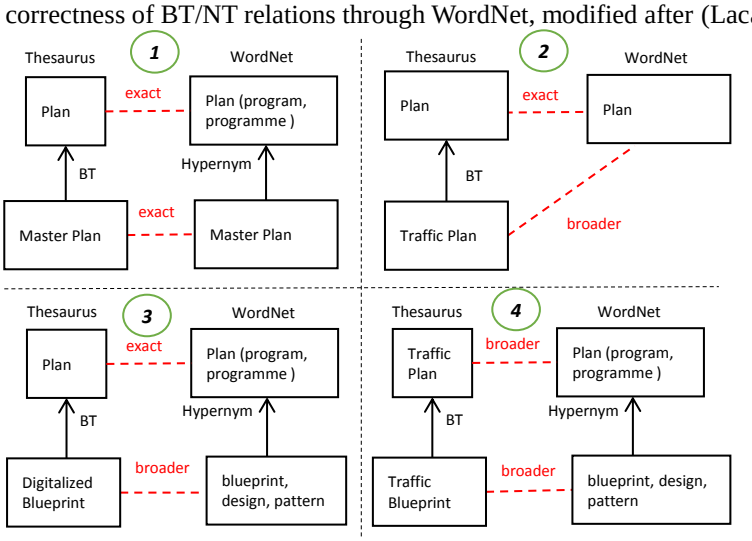
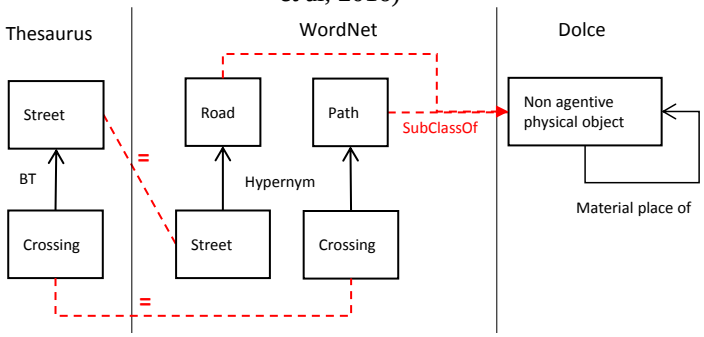


Figure 1 shows the different types of matching after the alignment between a thesaurus and Wordnet, and the consequences for the validation of BT/NT relations. The thesaurus-WorldNet alignment is mainly based on the string matching of labels, and in case of finding polysemy in Wordnet labels, the matching algorithm takes into account the multilingual features of the original thesaurus (the intersection of matched synsets in different languages may be unique) or the context provided by other thesaurus concepts in the branch with a monosemic matching in Wordnet. The hypernym relations in WordNet can validate the correctness of the BT/NT relations in the following cases: broader and narrower terms have an exact matching with hypernym and hyponym synsets (1); the narrower term has a broader matching with the hyponym synset (2); or the narrower term has a broader matching with hypernym synset (3). Otherwise, when both broader and narrower terms have a loosely matching with hypernyms and hyponyms (4), nothing can be inferred.

Indeed, the frequency of matchings similar to case 4 in figure 1 are quite common because WordNet only covers general concepts and many thesaurus concepts have broader matches with WordNet synsets (which may be even not connected at all). For validating those cases the tool proposes the use of an existent alignment between WordNet and DOLCE (Gangemi et al, 2003), which matches the upper levels of WordNet as subclasses of DOLCE. After this second alignment, if a BT/NT relation in the original thesaurus can be mapped with a DOLCE subclass, participation or location relation, this BT/NT relation will be classified as correct because it has a subordinate meaning compatible with DOLCE. Figure 2 shows an example of the validation of a BT/NT relation between *Street* and *Crossing* concepts. Through the alignment with WordNet, they are identified as hyponyms of *Road* and *Path*, which are in turn subclasses of a *physical object* in DOLCE. Then, as most of the relations between two *physical objects* are a kind of *location* relation, the tool validates the original BT/NT relation as correct.

Figure 2: Semantic correctness of BT/NT relations through WordNet and DOLCE, modified after (Lacasta et al, 2016)



3 Analysis of thesauri used in geographic metadata

Although there are not specific recommendations about the thesauri that should be used for describing spatial datasets or series, projects and technical guidelines related to the INSPIRE initiative (Zarazaga-Soria et al, 2007; Joint Research Centre, 2013) usually refer to thesauri such as GEMET, the Multilingual Agricultural Thesaurus - AGROVOC (Lauser et al, 2006) or the European Vocabulary Thesaurus - EUROVOC (European Union, 2017) due to several reasons: they have been created by public institutions (European Environment Agency, Food and Agriculture Organization, and European Parliament) and are publicly accessible; they are available in different languages and in standardized formats such as SKOS; and they provide a vocabulary broad enough to address the topics covered by different spatial datasets and series. Additionally, after analyzing the contents of the metadata catalogue of IDEE in September 2016, we discovered that 56% of 3,640 records containing a reference to a thesaurus concept different from INSPIRE Spatial Theme were making in most cases a reference to either GEMET, EUROVOC or AGROVOC. Whereas 40% of the total number of records were making a reference to any of these 3 thesauri, the use of other miscellaneous thesauri was not so relevant.

Therefore, we have analyzed these three thesauri, together with the general list of spatial data themes proposed in the INSPIRE directive. In particular, we have considered the concepts of these 4 thesauri in English, French, and Spanish. Table 1 shows the main characteristics of the thesauri: the number of concepts, the number of BT/NT relations, and the number of RT relations. Additionally, the table shows the percentage of records at IDEE metadata holdings containing a term from each thesaurus.

The results obtained after making the analysis of quality are shown in table 2. Although in general the quality of GEMET, AGROVOC and EUROVOC is high, there are some issues to be noted. In the case of GEMET it should be taken into account that definitions of concepts are only provided in English, there are some inconsistencies (11%) in the use of plural and singular forms of concept labels, and an estimated 24% of BT/NT relations are not semantically correct.

AGROVOC has also the problem of missing definitions. In fact, there are not definitions for concepts in any language. Additionally, 29% of RT relations are not informative as they connect concepts already in the same BT/NT hierarchy. With respect to BT/NT relations, it must be stressed that only 42% of relations could be verified by the tool because of the specificity of many terms, which could not be matched with

Table 2: Report of thesauri quality

Measure Name	GEMET	AGROVOC	EUROVOC	INSPIRE Spatial Themes
Property completeness analysis				
Completeness of preferred labels	99,83%	97.21%	97.49%	100.00%
Completeness of definitions	31,20%	0.00%	0.00%	100.00%
Property content analysis				
Non-existence of non-alphabetic characters in labels	92,61%	97.74%	91.18%	97.06%
Non-use of adverbs/initial articles in labels	99,56%	99.41%	96.36%	100.00%
Non-use of acronyms in preferred labels	99,00%	99.72%	96.32%	100.00%
Property context analysis				
Non-existence of duplicated labels	98,25%	99.25%	99.60%	99.02%
Consistent use of uppercase in labels	97,61%	96.44%	84.91%	100.00%
Consistent use of plurals in labels	88,50%	96.37%	82.74%	72.55%
Property complexity analysis				
Non-use of prepositional phrases in labels (en)	94,57%	99.31%	88.18%	99.02%
Non-use of too long noun phrases/conjunctions	93,53%	98.09%	70.03%	77.45%
Relation coherence analysis				
Informative RTs	96,84%	70.69%	100.00%	(*)
Completeness of BT/NT	100,00%	100.00%	100.00%	(*)
Non-existence of BT/NT cycles	100,00%	100.00%	100.00%	(*)
Semantic correctness of BT/NT	70,65%	32.97% (**)	69.75%	(*)

(*) Not applicable

(**) Only 42% of BT/NT relations could be verified by the tool

Wordnet synsets. This is the reason why only 33% of relations could be automatically checked as correct.

In the case of EUROVOC we have again the problem of lacking definitions, and inconsistencies in the use of singular and plural forms (17%). Additionally, it must be noted that there are also inconsistencies in the use of uppercases (15%), and a relevant amount of label texts seem to be too complex because they contain either prepositional phrases (11%) or very long noun phrases and conjunctions (29%). With respect to BT/NT relations, 30% of these relations are estimated as incorrect.

Last, it must be noted that the quality of the list of INSPIRE Spatial Themes, considered as a thesaurus, does not pose very specific problems but of course, it is not comparable with the rest of proper thesauri. Nevertheless, it can be observed that there are inconsistencies in the use of singular/plural forms (27%), and 22% of labels are considered too complex.

4 Conclusions

This work has shown how the tool for the quality analysis proposed in (Lacasta et al, 2016) can be customized to analyze the quality of thesauri most frequently used in geographic metadata. In particular, we have verified that the quality of GEMET, AGROVOC and EUROVOC is high in general terms, but there is still place for some improvements in these controlled vocabularies. On the one hand, a special effort should be devoted to provide definitions in different languages, and to check the consistency of the uniformity and simplicity of labels. Otherwise, metadata cataloguers may have problems in understanding the correct meaning of some concepts, and this refrains them to select these concepts for describing the theme of a resource. On the other hand, RT and BT/NT relations should be revised to assure that are informative and semantically correct respectively. Usually, metadata cataloguers browse through thesaurus relations to explore new concepts. If these relations are wrong, this may lead to keep some concepts hidden.

Once these thesauri are revised to correct these issues, our future work should be devoted to check if these improvements have a direct influence in the increase of quality of metadata records, and more particularly in the increase of the use of these controlled vocabularies. Notwithstanding this, an improved Human Computer Interface in cataloguing tools would probably also influence positively in the inclusion of more thesaurus concepts. Nowadays, the functionality of browsing thesauri during the metadata creation process is very limited, and this affects negatively in the usability of thesauri. Additionally, new developments in cataloguing tools should consider the applicability of text mining techniques to recommend automatically controlled vocabularies. The names of tables, columns and text values contained in datasets could provide input data for automatic topic categorization. Another possibility could be to provide mappings between mandatory vocabularies such as the INSPIRE Spatial Data Themes and other more specific vocabularies.

Acknowledgements

This work has been partially supported by Keystone COST Action IC1302 and the National Geographic Institute (IGN) of Spain. Additionally, the work of Dayany Díaz-Corona has been supported by a grant from Universidad de Zaragoza and Banco Santander.

References

Albertoni, R., De Martino, M. and Quarati, A. (2016). Integrated Quality Assessment of Linked Thesauri for the Environment. In: *Proceeding of 5th International Conference on Electronic Government and the Information Systems Perspective (EGOVIS)*, Porto, Portugal, 2016, pp. 221-235.

European Commission (2008). COMMISSION REGULATION (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata. Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX%3A32008R1205%3AEN%3ANOT [Accessed 2nd February 2017].

European Environment Agency (2017). GEneral Multilingual Environmental Thesaurus (GEMET). Available from: http://www.eionet.europa.eu/gemet [Accessed 2nd February 2017].

European Union (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX%3A32007L0002%3AEN%3ANOT [Accessed 2nd February 2017].

European Union (2017). EuroVoc, Multilingual Thesaurus of the European Union. Available from: http://eurovoc.europa.eu [Accessed 2nd February 2017].

Fellbaum, C. (ed) (1998). WordNet: An Electronic Lexical Database, MIT Press.

Fugazza C. and Luraschi, G. (2012). Semantics-Aware Indexing of Geospatial Resources Based on Multilingual Thesauri: Methodology and Preliminary Results. *International Journal of Spatial Data Infrastructures Research*, 7, 16-37.

Gangemi, A., Guarino, N., Masolo, C. and Oltramari, A. (2003). Sweetening WORDNET with DOLCE, *AI Magazine*, 24(3), 13-24.

International Organization for Standardization (1994). ISO 8402:1994 - Quality management and quality assurance.

International Organization for Standardization (2011). ISO 25694:2011 - Thesauri and Interoperability with other Vocabularies.

Joint Research Centre (2013). INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119. Available from: http://inspire.ec.europa.eu/file/1557/download?token=UaQBcRvQ [Accessed 2nd February 2017].

Lacasta, J., Falquet, G., Zarazaga-Soria, F.J. and Nogueras-Iso, J. (2016). An automatic method for reporting the quality of thesauri. *Data & Knowledge Engineering*, 104, 1-14.

Lauser, B., Sini, M., Salokhe, G., Keizer, J. and Katz, S. (2006). Agrovoc Web Services: Improved, real-time access to an agricultural thesaurus, *Quarterly Bulletin of the International Association of Agricultural Information Specialists (IAALD)*, 1019-9926 (2), 79-81.

Suominen, O. and Mader, C. (2014). Assessing and Improving the Quality of SKOS Vocabularies, *Journal on Data Semantics*, 3(1), 47-73.

Zarazaga-Soria, F.J., Nogueras-Iso, J., Latre, M.Á., Rodriguez-Pascual, A., López, E., Vivas, P. and Muro-Medrano, P.R. (2007). Providing Spatial Data Infrastructure Services in a Cross-Border Scenario: SDIGER Project. In: *Research and Theory in Advancing Spatial Data Infrastructure Concepts*. ESRI Press, pp. 107-119.