Automating the Thematic Characterization of Geographic Resource Collections by Means of Topic Maps

J. Lacasta, J. Nogueras-Iso, R. Tolosana, F.J. Lopez, F.J. Zarazaga-Soria
Computer Science and Systems Engineering Department, University of Zaragoza
María de Luna, 1. 50018-Zaragoza (Spain)
{jlacasta, jnog, rafaelt, fjlopez, javy}@unizar.es

SUMMARY

Spatial Data Infrastructures at national or higher levels usually comprise the access to multiple geographic data catalogs. However, with classical search systems it is difficult to have a clear idea of the information contained in the metadata holdings of these catalogs. Topic maps are representation systems that allow the access to data with emphasis in the find-ability. This paper describes a process to generate a hierarchical topic map from a collection of geographic metadata records and the uses that the generated topic map can have in the context of a spatial data infrastructure to facilitate the access to the data.

INTRODUCTION

Spatial Data Infrastructures (SDI) provide the framework for the optimization of the creation, maintenance and distribution of geographic information at different organization levels (e.g., regional, national, or global level) and involving both public and private institutions (Nebert, 2001). Geographic data catalogs are one of the main components of an SDI as they provide services for searching geographic information by means of their metadata and according to a particular search criterion. But from the perspective of a European or Global SDI, the problem is that usually there is not a unique geographic data catalog. Quite the opposite, hundreds of geographic data catalogs should contribute from the different nodes of the SDI. In such situation, where searches are distributed to different metadata catalog, it is difficult to have a clear idea of the information described in each metadata collection. Therefore, it would be interesting to provide the mechanisms that give a general overview of the contents of a geographic catalog (collection). Furthermore, this is even useful in the case of having individual geographic data catalogs in order to facilitate a summary of the contents for those users accessing for the first time.

The objective of this work is to describe a way to extract a summary of the contents of a metadata collection. This summary of contents will be focused on the thematic description of the collection, that is, a topic map. Topic maps are a representation of knowledge, with an emphasis on the find-ability of information. In fact, there is an international standard ISO1325 (ISO, 2003c) that has already defined a similar topic map concept (the standard proposes a model for representation and defines the interchange format). Topic maps allow easy and selective navigation to the requested information, showing a thematic view of the collection of metadata, facilitating in that way the access to the information. The main objective of the generated topic map will be to provide a thematic global vision of the collection giving information about what metadata records in the collection use each term of the topic map and the weight of each keyword in the collection. Another objective is to obtain a classification of the collection more reduced than the list of all topics used in the metadata records, to circumscribe the collection to a concrete theme.

Among the contributions in the line of dividing metadata collections into clusters of resources with similar characteristics making emphasis in the graphical visualization of these groups it can be highlighted the following. (Boulos et al., 2001) presents a system that creates a graphic topic map to enhance the access to a medical database using a graphical representation to locate the different topics over a graph of the human body. (Krowne and Halbert, 2004) shows the effectiveness of different

clustering techniques for heterogeneous collections of metadata records according to different factors, as the time used to locate a resource, the number of clicks needed to reach it or the number of failures in locating the resource. (Schlieder et al., 2001; Schlieder and Vogele, 2002) insist on the idea of accessing to metadata collections through a network of intelligent thumbnails, being those thumbnails either concepts or locations. For instance, in the case of locations, the network of thumbnails corresponds to the hierarchical structure of administrative toponyms. (Demšar, 2004) and (Podolak and Demšar, 2004) provide a visual data mining approach to explore in an easier way the complex structure/syntax of geographical metadata records. (Podolak and Demšar, 2004) also describe a system to clusterize a collection of metadata into clusters of similar metadata elements using the cluster algorithm of (Fisher, 1987). (Albertoni et. al., 2003) describes a system to visually show traditional representations of statistical information about a collection to the user to facilitate him the identification of patterns.

The contribution of this paper is to show the benefits of a hierarchical topic map as a clustering and visualizing tool to locate resources in a SDI, and the method to generate it automatically on the basis of the thesaurus structure of the keywords used in the metadata collection. This paper is organized as follows. First, the method for the extraction of the topic map is shown. Second, a summary of the preliminary experiments is done. Third, the uses of a topic map in a SDI are described. The paper is finished with some conclusions and description of future work.

METHOD FOR THE EXTRACTION OF THE TOPIC MAP

This section describes the steps to extract a topic map from a metadata collection using the keywords section of the metadata records. At this step of the development we have simplified the problem to a collection of metadata records that, in their keywords section, use terms from a unique thesaurus, leaving for future work the analysis of the problematic of mixing several thesauri into a homogeneous topic map. The process to automatically generate the topic map from the metadata collection is shown in detail in Figure 1.

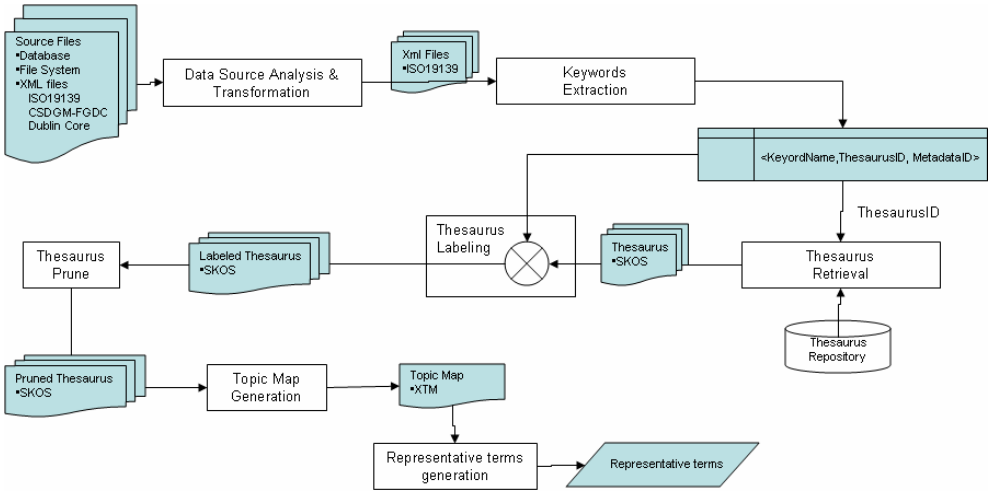

Figure 1: Topic Map generation Process

The first step is to obtain the metadata collection in a format readable for our system. The metadata records can be stored in different repository sources as a metadata catalog, a relational database, or even in a directory of XML files; and their structure can follow different standards as

(ISO, 2003a), (ISO, 2003b) or (FGDC, 1998). So, the first step of the process is the identification of the source format and the transformation of the metadata records into a list of XML files following the ISO-19139 (ISO, 2005) standard.

The second step is the analysis of the metadata collection and the generation of the list of triplets <KeywordName, ThesaurusID, MetadataID> with the values of the keywords of the metadata, the identifier of their source thesaurus and the identifier of the metadata record where the keyword has been found. Such triplets are the base elements used to construct the topic map classification.

The third step consists in the retrieval of the thesaurus used to create the keywords of the metadata collection. A thesaurus can be provided by different kind of applications in different formats, but to be able to use it in our system to create a topic map it has to be manually obtained and transformed into SKOS core format. SKOS core is an RDF vocabulary for expressing the basic structure and content of concept schemes as thesauri, classification schemes, subject heading lists, taxonomies, terminologies, glossaries and other types of controlled vocabulary (Alistair et. al., 2005). This format has been selected as import format since it is becoming in de facto standard for the interchange of thesauri as many thesauri covering different areas are being generated in this format.

Once the thesaurus used in the metadata records has been identified and retrieved, it is labelled with the information of the triplets obtained in the second step. This labelling process considers not only the direct uses but also the inherited through the thesaurus hierarchy, storing separately the direct relations from the inherited ones. This separation allows identifying in future steps if a term is used directly by some metadata records or if the used is one of its descendents.

The next step consists in the prune of the branches and leaves of the labelled thesaurus whose terms are not referenced directly of by inheritance in any metadata record of the collection. Only terms with no direct or inherited references added in the previous step are deleted. Inner nodes with inherited relations are not deleted because for the use in a search system is better if the user have more detailed hierarchy to select terms for a query. It is useful when the user is not an expert in the themes of the metadata collection, because if he does not understand the meaning of some of the more specific terms used in the lower levels of the hierarchy, he can select one of its more general ancestors to retrieve the associated information.

The labelled thesaurus can then be provided in an extended SKOS format to a search service, to use it as topic map to navigate through the data or as part of a keyword selection system, or in a retrieval system to provide information about how many results are going to be returned when a keyword is selected. An example of the SKOS representation of an element of the topic map can be seen in Figure 2. The extension done to the basic SKOS format has been to add two new properties “directReference” and “inheritedReference” used to store the identifiers of the metadata records that use directly of by inheritance the keyword.

```
<skos:Concept rdf:about="http://www.eionet.eu.int/gemet/concept/2405">
   <skos:prefLabel xml:lang="es">ciencias de la tierra</skos:prefLabel>
   <skos:prefLabel xml:lang="fr">sciences de la terre</skos:prefLabel>
   <skos:definition xml:lang="en">The science that deals with the earth or any part thereof; includes the disciplines of geology,
      geography, oceanography and meteorology, among others. (Source: MGH)</skos:definition>
   <topicMap:directReference>digital-data-30-boundary<topicMap:directReference>
   <topicMap:directReference> digital-data-30-P-13-cells<topicMap:directReference>
   …
   <topicMap:inheritedReference>digital-data-30-P-4-conventional<topicMap: inheritedReference>
   …
   <skos:narrower rdf:resource="http://www.eionet.eu.int/gemet/concept/5270"/>
    …
</skos:Concept>
```

Figure 2: Labelled SKOS Concept

This generated file can be directly used as topic map if the system to use it is able to read it, or it can be transformed into a standard format, to allow that many existent applications could directly use this topic map. Among the available formats for topic maps representation, the XTM format (Pepper, 2001) seems to be the most adequate for its use as interchange format for the topic map generated, by its simplicity and by the number of applications able to read and visualize it.

The last step consists in the extraction of the main themes of the collection by means of analyzing the topic map generated in the previous step. This can be manually done using visualization tools that allow the load of SKOS or RDF (as Protégé[11]) and detecting manually the most used themes, but the updates that usually are done on the metadata collections can change the main themes of the collection, so to avoid the human work, an automatic process to extract this information has been designed. The process described here is based in the idea of concept density used for the disambiguation of free text in (Agirre and Rigau, 1996). The objective is to obtain the representative nodes of the tree that aggregate a relevant percentage of records in the metadata collection.

Clustering techniques analyze different properties of data to statistically group together the most similar. Elements in a cluster share common features according to a similarity criteria (Demšar, 2004). Inherently, the topic maps described above can be considered as hierarchical clusters. However, our intention in this last step is to obtain a 1-dimensional cluster that summarizes the collection at a first glance. Here, the hierarchical structure of the thesaurus provides a thematic context of similarity that enables terms of the same branch to be summarized by the root node of the branch. In order to identify this 1-dimensional cluster the Formula 1 has been proposed.

$$\frac{\sum_{\forall nodeInBranch} \text{numberOfRecords}}{\sum_{\forall recordsInCollection} \text{numKeywords}} > \text{threshold}$$

Formula 1: Topic map clustering formula

The criterion used to identify a representative node divides the number of records containing this node (or any of its descendants) by the number of keywords in the collection. If this value is greater than a threshold, then this node is considered as a relevant node. At present this threshold has been selected experimentally, using a range of values between 0.05 and 0.2 (see following section).

EXPERIMENTS IN THE CREATION OF HIERARCHICAL TOPIC MAPS

This section shows a summary of the experiments made to analyse the viability of the process for the automatic creation of a hierarchical topic map from the keywords of a collection of metadata. As metadata corpus for experiment, the contents of the Geoscience Data Catalog[12] at the U.S. Geological Survey (USGS) were downloaded. The USGS is the science agency for the U.S. Department of the Interior that provides information about Earth, its natural and living resources, natural hazards, and the environment. Despite being a national agency, it is also sought out by thousands of partners and customers around the world for its natural science expertise and its vast earth and biological data holdings. This metadata collection was processed as indicated in (Nogueras et al, 2005) until a collection of 753 metadata records compliant with (CSDGM, 1998) standard was obtained. The 626 keywords in the metadata collection, obtained from the GEMET[13] thesaurus have been the selected to

[11] http://protege.stanford.edu/
[12] http://geo-nsdi.er.usgs.gov/
[13] http://www.eionet.eu.int/GEMET

be used in the experiment. The GEMET version used contains 5542 terms from which the collection of metadata uses 104 different ones, which is about 1.9% of the thesaurus size.

The topic map creation process has been applied to this collection using GEMET as base for the creation of the topic map. The topic map generated contains 216 nodes (the 104 used in the collection plus 112 inner nodes) reducing the percentage of GEMET to the 4% of its size. This huge reduction of size will provide a user a much more adjusted selection tool to locate a resource.

To visualize the correctness of the topic map generated and to provide quickly to the users a tool able to navigate by the topic map and to locate the associated information, we have selected the TMNAV tool created in the TM4J project (TM4J, 2001). This tool allows the visualization of topic maps stored in XTM format (Pepper, 2001) providing a graphical visualization of the properties of the records and the navigation by the relations. A branch of the generated topic map is shown in Figure 3 as example. In this example is shown the node *Atmosphere (air, climate)* of the topic map, it can be seen its relations with other terms of the topic map, as *climate* and the two metadata records of the collection that indirectly contain this term of the topic map (One of their children has some occurrences in the metadata records of the collection). The example shown has not direct relations (metadata records that directly contain the term of the topic map) but if they were, they would have been also shown in the visual representation.

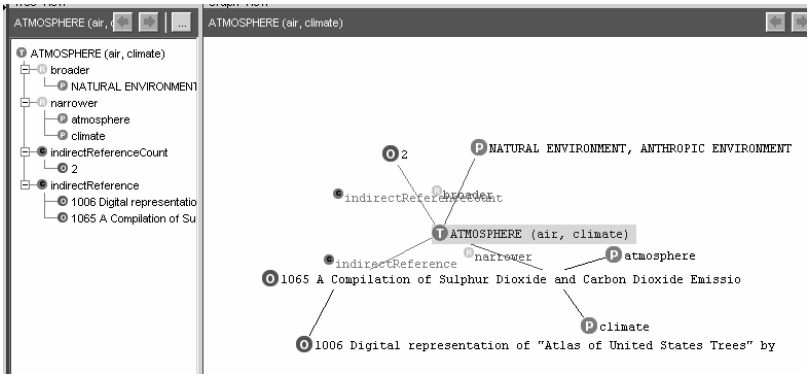

Figure 3: Structure of a branch of the topic map

When the topic map is presented to the user in a graphical view, at first sight he can see how many results is going to obtain if he selects a term from the hierarchy (results that contain the selected keyword or one of its descendent). Then, the user can refine the selection navigating for the tree, until it finds the most adequate term to query. This avoid him the execution of several queries that do not restrict the collection enough and produce too many results or select terms of a thesaurus not used in the metadata records and produce zero results.

Once the topic map was generated, the next step of the experiment has been to extract automatically the main theme of the collection. The algorithm for the selection of relevant nodes (shown in previous section) was applied to the topic map using as threshold the values of 0.05, 0.1 and 0.2. With these parameters the thematic classifications of Figure 4 was obtained.

Threshold = 0.05	Threshold = 0.1	Threshold = 0.2
human activities and products, effects on the environment	human activities and products, effects on the environment chemistry	**human activities and products, effects on the environment**
chemistry, substances, processes	chemistry, substances, processes	-----
products, materials	-----	-----

product	-----	-----
fuel	-----	-----
fossil fuel	-----	-----
coal	-----	-----
material	material	-----
raw material	raw material	-----
natural gas	**natural gas**	-----
social aspects, environmental policy measures	social aspects, environmental policy measures	social aspects, environmental policy measures
research, sciences	research, sciences	research, sciences
science	science	science
natural science	natural science	natural science
earth science	earth science	**earth science**
geology	**geology**	-----
marine geology	-----	-----
natural environment, anthropic environment	-----	-----

Figure 4: Threshold effect for the extraction of representative nodes

Figure shows the keywords selected as main themes of the metadata collection given a threshold. From the terms selected during the generation process, only those that has no descendants in the extracted terms are marked in bold face as main keywords, the rest of the selected terms are their hierarchical ancestors in the topic map. This example shows that the main themes of the metadata collection are “earth science” and “human activities and products, effects on the environment” terms, given that they have been obtained with the highest threshold. It is also shown that the secondary themes are “geology” (child of “earth science”) and “natural gas”, both obtained if the clustering threshold is reduced to 0.1. The terms “coal”, “marine geology” and “natural environment, anthropic environment**”** are less relevant, given that they have been obtained with the lowest threshold.

USES OF TOPIC MAPS IN AN SDI

The creation of a topic map from a metadata collection provides several benefits to an SDI as concerns the improvement of retrieval systems, graphical search systems, navigation for the collection of metadata records and analysis of the metadata collection. A detailed description of the benefits could be categorized as follows:

- Contribute to the improvement of distributed catalog strategies. The Open Geospatial Consortium (OGC) specifies in (Whiteside, 2005) the basic metadata that every service of a SDI should return. Between those metadata, the keywords section can indicate the theme of the data provided by the service. Those keywords can be used by a distributed catalog to redirect user queries only to the local catalogs with the appropriate themes, reducing in that way the response time of the query and the network overhead. These metadata can be created manually for the system administrator analyzing the metadata collection with visual data mining tools but with the inconvenient of having to do the same again each time there is an update. In this situation, a hierarchical topic map can be used as base for the automatic creation of the keywords of the collection, facilitating their update if the collection changes.
- Facilitate the navigation. Topic maps provide a different way to navigate for the data in the collection. Looking at the topic map, you can directly locate the list of metadata records that use a term of the topic map, avoiding free text or controlled queries that can easily produce empty result sets. An example of how a graphic topic map can be used to facilitate navigation is the Health Cyber Map project proposed in (Boulos et. al., 2001).
- Facilitate the construction of queries. In metadata creation process it is usual to provide controlled lists or thesauri to facilitate the creation of metadata records and to use the same structures in search systems to facilitate the location of resources. Providing directly to the user the thesauri used to create the keywords section of the metadata records sometimes is

not useful because the thesauri can be too many and/or too big and, for many possible selections, the query system can produce zero results. Here, the automatic creation of a hierarchical topic map where the number of associated metadata record is shown can replace the thesauri for selection of terms, given that it facilitates the user to guess if the collection contains useful data and reduces the possibility of constructing queries that produce zero results.

- Metadata quality evaluation. The identification of the main themes of a collection can be used to analyze if the thesaurus used in the metadata collection is the most adequate. For example, if it is detected that almost all the keywords of the collection are in the hydrology path of UNESCO[14] thesaurus, possibly, to create the metadata records, instead of UNESCO, a thesaurus specialized in hydrology should have been used to allow the selection of more specialised keywords in the metadata records, to provide to the user a more complete description of the data.

CONCLUSIONS AND FUTURE LINES

This paper has shown how to extract a topic map from a collection of geographic metadata records. The method proposed assumes that the terms used in the keywords section have been selected from a well established thesaurus, and the aim of the method is to extract a hierarchical topic map that reduces significantly the size of the original thesaurus. Additionally, the method proposed also suggests a formula to obtain an even more reduced set of representative nodes (a 1-dimensional cluster), which summarizes the main themes of the collection at a first glance.

Besides, the paper details the benefits of the use of topic maps for the improvement of the services offered by a Spatial Data Infrastructure. The generated topic maps may contribute to the improvement of distributed catalog strategies, facilitate the navigation, the construction of queries and can help to analyze the quality of the metadata contents. Overall, it can be highlighted that the topic maps proposed facilitate enormously selection of the terms in a query system and that the generated themes (the reduced set of representative nodes) gives a synthesized and accurate summary of the contents of the metadata collection.

As future lines, it can be stated that the problem of creating a topic map has been simplified here by adding the restriction that the metadata collection only uses a thesaurus in the keywords section of each metadata record. Future work will eliminate this restriction and it will analyze the problem of having, in the metadata records, keywords from different thesauri to identify the problems produced when is needed to mix different thesauri into one. Another area of work is the creation of a topic map from thematic keywords not obtained from a thesaurus but from plain lists or free text and the creation of topic maps from other elements of the metadata structure. Finally, the formula proposed for the creation of a 1-dimensional thematic cluster must be still improved. On the one hand, the present algorithm lacks the ability to automatically detect the most appropriate clustering threshold to generate the main themes of the collection (i.e. the threshold has to be selected experimentally). And on the other hand, other alternatives for thematic clustering should be analyzed and compared with the approach proposed here.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Ministry of Education and Science through the project TIC2003-09365-C02-01 from the National Plan for Scientific Research, Development and

[14] http://www.ulcc.ac.uk/unesco/

Technology Innovation. The work of J. Lacasta has been partially supported by a grant (ref. B139/2003) from the Aragon Government.

BIBLIOGRAPHY

Agirre E., Rigau G., 1996. Word Sense Disambiguation Using Conceptual Density. In: Proceedings of the 16th International Conference on Computational Linguistics (Coling'96) Copenhagen, Denmark, 1996, pp. 16-22.

Albertoni R., Bertone A., Demšar U., De Martino M., Hauska H., 2003. Knowledge Extraction by Visual Data Mining of Metadata in Site Planning. In: Proceedings of the 9th Scandinavian Research Conference on Geographic Information Science, ScanGIS'2003, Espoo, Finland, pp. 119-130.

Alistair M., Matthews B., Wilson M., 2005. SKOS Core: Simple Knowledge organization for the WEB. In: Proceedings of the International Conference on Dublin Core and Metadata Applications. September 2005. Madrid, Spain.

Boulos MN., Roudsari AV., Carson ER., 2001. Towards a Semantic Medical Web: HealthCyberMaps Dublin Core Ontology in Protégé-2000. In: Fifth International Protégé Workshop, SCHIN, Newcastle, UK.

Federal Geographic Data Committee (FGDC), 1998. Content Standard for Digital Geospatial Metadata, version 2.0. Document FGDC-STD-001-1998, Metadata Ad Hoc Working Group.

Demšar, U., 2004. A visualization of a Hierarchical Structure in Geographical metadata. In: Proceedings of the 7th AGILE Conference on Geographic Information Science, 29 April - 1 May 2004, Heraklion, Greece.

Fisher DH., 1987. Knowledge Acquisition Via Incremental Conceptual Clustering. Machine Learning 2, 139-172.

ISO, 2003a. Geographic information – Metadata, ISO19115:2003. International Organization for Standardization (ISO).

ISO, 2003b. Information and documentation - The Dublin Core metadata element set. ISO 15836:2003. International Organization for Standardization (ISO).

ISO, 2003c. Information technology -- SGML applications -- Topic maps. ISO13250:2003 International Organization for Standardisation (ISO).

ISO, 2005. Geographic information -- Metadata -- XML schema implementation Draft. ISO/WD 19139, ISO/TC 211. International Organization for Standardization (ISO).

Krowne A., Halbert M., 2004. An Evaluation of Clustering and Automatic Classification For Digital Library Browse Ontologies. Metacombine Project Report. http://metacombine.org

Nebert, D., (ed). 2001. Developing Spatial Data Infrastructures: The SDI Cookbook v.1.1. Global Spatial Data Infrastructure, May 2001. http://www.gsdi.org.

Nebert D., Whiteside A. (eds), 2004. OpenGIS Catalog Services Specification, v2.0. OGC 04-021r2. Open Geospatial Consortium.

Nogueras-Iso J, Zarazaga-Soria FJ, Muro-Medrano PR., 2005. Geographic Information Metadata for Spatial Data Infrastructures - Resources, Interoperability and Information Retrieval. Springer Verlag.

Pepper S., Moore G. (Eds.), 2001. XML Topic Maps (XTM) 1.0. TopicMaps.Org.

Podolak I., Demšar U., 2004. Discovering structure in geographical metadata. In: Proceedings of the 12th conference in Geoinformatics. June 2004.Galve, Sweden.

Schlieder C., Vögele T., Visser U., 2001. Qualitative Spatial Representation for Information Retrieval by Gazetteers. In: Proceedings of Conference of Spatial Information Theory COSIT, Vol. 2205, Morrow Bay, CA, pp. 336-351.

Schlieder C., Vögele T., 2002. Indexing and Browsing Digital Maps with Intelligent Thumbnails. In: Proceedings of Spatial Data Handling 2002 (SDH'02), Ottawa, Canada

TM4J, 2001. TM4J project team. Homepage of the TM4J project. 2001. http://tm4j.org/.

Whiteside A. (ed), 2005. OGC Web Service Common Specification, v1.0. OGC 05-008. Open Geospatial Consortium.