Using a Semantic Approach for a Cataloguing Service

Paul Boisson, Stéphane Clerc, Jean-Christophe Desconnets, Thérèse Libourel Rouge

Using a semantic approach for a cataloguing service

Paul Boisson[1], Stéphane Clerc[1] Jean-Christophe Desconnets[1], and Thérèse Libourel[2]

[1] IRD (Institut de Recherche pour le Développement), 34394 Montpellier Cedex 5, France,
boisson@teledetection.fr, clerc@teledetection.fr, jcd@teledetection.fr,
[2] LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier), 34392 Montpellier Cedex 5, France,
libourel@lirmm.fr

Abstract. Environmental applications (support for territorial diagnostics, monitoring of practices, integrated management, etc.) have strengthened the case for efforts in the establishment of sharing and mutualisation infrastructures for georeferenced information. Within the framework of these initiatives, our work has led us to design and create a tool for cataloguing resources for environmental applications. This tool can be used to catalogue different types of resources (digital maps, vector layers, geographical databases, documents, etc.) by using the ISO 19115 standard, and offers a search engine for these resources. The goal of this proposition is to improve the relevance of search engines by relying on semantic knowledge (thematic and spatial) of the concerned domains. In the first stage, the proposition consists of helping the user in his search by offering mechanisms to expand on or to filter his query. In the second stage, we use the results obtained and the underlying semantics for a global presentation of the results.

1 Introduction

Environmental applications (support for territorial diagnostics, monitoring of practices, integrated management, etc.) have strengthened the case for efforts in the establishment of sharing and mutualisation infrastructures for georeferenced information[3]. Within the framework of these initiatives, our work has led us to design and create a tool for cataloguing resources, a fundamental and indispensable requirement for data infrastructure. The goal now is to make this tool's search engine more relevant by drawing on the semantics of the concerned domains.

Section 2 presents other works on the same subject. Section 3 of this article presents the context within which our work is placed and lists our motivations. In section 4, we recall the main features of the MDweb cataloguing tool by describing its original characteristics (genericity, use of a thesaurus and a reference

[3] http://inspire.jrc.it/

base of geographical objects to control thematic and spatial descriptors). In fact, it is by the use of thematic and spatial reference bases that we propose to improve the relevance of the tool. In sections 5 and 6 we present the basic approach of our work. It involves the concept of expansion of a search based on spatial and thematic semantics and on the distribution of semantic relationships contained in the resources found. Section 7 discusses the future perspectives opened up by this work.

2 State of the art

Since our intention was to improve the relevance of the MDweb resources search engine (see section 3), both at the level of the search itself and in the exploitation of its results, we looked for similar approaches in existing work done in this field.

Some cataloguing tools based on the use of metadata standards in the geographic information domain currently exist, for example, GéoConnexions (Canadian geospatial data infrastructure), Geospatial Metadata tools[4], M3Cat[5], EU Geo-Portal, GeoNetwork (FAO), GEoNorge, etc.

Some cataloguing services with a semantic-Web approach, in any field, have already enhanced their functionalities by drawing on the use of thesauri or ontologies. We can mention, for example, the French project Cismef (Catalogue and Index of French-language Medical Sites) [9] which has developed an intelligent system for the search of information based on a set of metadata elements describing resources in the medical domain. This system uses a medical ontology that makes inferences to provide results that are more user relevant. In this framework, ontology offers the possibility to the underlying search engine of automatizing reasoning, on the basis of the user's initial query, with the aim of providing better results. The mechanism consists of reformulating the initial query so that it returns additional results or, on the other hand, of filtering it to retain only the most relevant results. The use of ontologies depends on their type. In the context of geographic information, the process of query reformulation can be based on the thematic aspect by using the relationships between the search keywords (as in the Cismef project [9] or even [3]). We can also draw on the spatial aspect for using spatial relationships, as is done in the European project SPIRIT [4]. Other improvements in the relevance of information search results have been proposed; they rely on techniques based on the statistical cooccurrence of terms or even on the use of the user's profile [2].

The semantic use of the results depends on the formalisms proposed for knowledge representation. Even though there already exist formalisms such as those of conceptual graphs of Sowa [10], or RDF of Guha (1987), or Topics Maps, based on research by the Davenport Group in the 1990s, applications that allow graphical visualization of knowledge are few and far between. The OKS[6] (Ontopia Knowl-

[4] http://www.fgdc.gov/metadata/geospatial-metadata-tools
[5] http://www.intelec.ca/francais.html
[6] http://www.ontopia.net/

edge Suite) project of Ontopia includes several ontology management modules using Topic Maps formalisms, of which one (Omnigator[7]) graphically displays the Topic Map. Similarly, the KAON[8] (Karlsruhe Ontology Management Infrastructure) programme also manages ontologies by using its own formalism. It allows graphical viewing of the ontology on which one is working. Amongst information-search tools, most present results traditionally: the query results are in the form of a list. We can, however, note some efforts towards visual representation, such as the Kartoo[9] metasearch engine which displays the results in cartographic form. Information is displayed in the form of "semantic" graphs, nodes corresponding to resource results, with arcs connecting nearby nodes semantically.
Drawing inspiration from this body of work, we shall now explain, in the following sections, our ideas for improving the relevance of the MDweb tool.

3 Context

This research was conducted by a group of multidisciplinary scientists bringing together information-technology and thematics specialists. The varied applications and interests of these scientists have, as a common point, the mutualisation and sharing of data, processes and knowledge. Their combined efforts led to the design and creation of a cataloguing tool, MDweb, which has been used in several projects and activities linking diverse communities (Coastal zone integrated management Syscolag [1], National Environmental Information System of Cape Verde[10], Network of Long Term Ecological Observatories ROSELT[11] [5], Programme DeSurvey: a surveillance system for assessing and monitoring of Desertification[12]).
The tool was designed to provide the concerned communities with a way of managing their internal resources and to distribute, over the Web, the knowledge necessary to locate these resources. Incorporating the dimension of *searching for information on the Web* requires tackling the inherent problems associated with information searches on the Web. Existing search engines are no longer capable of easily locating the concerned resources because the traditional Web collects heterogeneous information that is, for the most part, unstructured and which can only be processed by humans. To alleviate this shortcoming, the proposed tool uses metadata to annotate resources. Since the resources are georeferenced, the annotation conforms to the ISO standard: ISO 19115. While the structuring proposed by this standard (as by all other metadata standards) is only a first step towards automatized searches, the use of domain semantics is still in its infancy. In the general context of the Web, the final stage will be, as proposed by

[7] http://www.ontopia.net/omnigator/models/index.jsp
[8] http://kaon.semanticweb.org/
[9] http://www.kartoo.com
[10] http://www.sia.cv/
[11] http://mdweb.roselt-oss.org/
[12] http://www.desurvey.net/

Tim Berners-Lee, the construction of a semantic web [11] whose contents will be comprehensible both by humans and machines, implying intelligent automatized searches of information.
The construction of ontologies thus becomes necessary to complement resource annotation. We have progressed in this direction by incorporating a shared representation of the knowledge domain. This will permit the annotation from a controlled vocabulary and it is on this explicit knowledge that we shall rely to enrich the search process and the results presentation.

4 The MDweb tool

4.1 Principles and genericity

Principles In view of the variety of possible applications in the diverse communities that use georeferenced information, MDweb was designed as a generic, multi-lingual, multi-standard tool for cataloguing and searching georeferenced information[13]. It can be used to create, manage and search one or more catalogues via the Web. MDweb is a server-side application. It has been designed so that it can be independently deployed on Windows or Linux operating systems. To ensure interoperability with other cataloguing applications, MDweb implements, most notably, the international standard for geographical information metadata, ISO 19115 [7], as far as metadata structuring is concerned, and the Catalog Service specifications of the OpenGIS Consortium [8] to ensure interoperability of the cataloguing service by implementing the protocol z3950-ISO 23950 [6].

Genericity The tool's genericity is based on its core, which is a generic database (metabase) whose relational schema can be broken up into four subschemas:

- S1: storage schema for the standard's dictionary,
- S2: storage schema for the metadata template structures and their parameters inherited from a standard,
- S3: storage schema for metadata,
- S4: storage schema for the contents of the user interfaces (labels, user parameters).

This core thus ensures (S1) the storage of elements and structures of diverse metadata standards (ISO 19115, FGDC, Dublin core). The use of the storage schema of the standard's dictionary allows, by calculation, to arrive at new, varied structures, adapted to the needs of a community (templates or metadata profile) without modifying the relational metadata storage schema (S3) and the user interfaces. The values stored in (S4) and (S2) allow the creation and customization of user interfaces as also of diverse parameters or values necessary for the entry of metadata.

[13] It was designed and developed by our multi-disciplinary group that included members from IRD, LIRMM and Cemagref.

4.2 Originality

Its genericity apart, the originality of the proposed service can be perceived from two different angles.
Firstly, conscious of the fact that headings included in metadata structures are insufficient for the annotation, we have complemented them by introducing the semantics of the concerned domains. Two reference bases have been added to the tool: a thematic reference base and a spatial reference base. The thematic reference base describes the semantics by the intermediary of explicit models. These models are described by an established and shareable vocabulary (the very basis of the concept of ontology) within the thesaurus.
The second characteristic, important in our view, and one that we have included in the tool, consists of helping the user in all his tasks. The input and search stages in a metadata service can very rapidly become restrictive and even disillusion users. MDweb therefore offers a set of user aids, the most important one being, as far as our work is concerned, the linking of the search engine's multi-criteria search interface with the semantic aspects of the two reference bases. Normally, a multi-criteria search is constructed by the combination of four criteria: **What** resource type? To **what subject** does the resource apply? **Where** is it located? and **When** was it created?
Nevertheless, as for any search engine, the response may be "silent" or, on the other hand, "too verbose". Our first aim therefore is to improve the search engine's behaviour in such situations. Subsequently, we hope that the results of any search can be best used (beyond a mere listing) by bringing out the implicit knowledge that they contain.

5 Improving searches

The first idea is to improve the relevance and number of results returned as a response to a user's query by relying on the semantic addition to the tool (cf. section 4). Towards this end, we propose to adapt the *query expansion* mechanism to our context. Query expansion can be defined as a process that modifies the initial user query to better respond to his query. If the system remains *silent* to the user's search, expansion will take the form of enrichment by widening the search scope. If, on the other hand, the system responds *verbosely*, i.e., it returns too many results, the process filters the query to refine the response to the user. Different automatic and interactive expansion mechanisms have thus been developed.

5.1 Thematic expansion

Thematic expansion of the search is based on a modification of the initial query (enrichment or filtering) using the thesaurus that the metadata service relies on. The search engine complements or refines the query using thesaurus keywords. These keywords are selected because they are in semantic relationships with

those of the user. We recall here that terms in the thesaurus are organized in a hierarchy and are connected between themselves by different relationships such as synonym links, associated-terms links, etc. (details of possible relationships are presented in Table 1).

BT: "Broader term" relationship
NT: "Narrower term" relationship
UF: "Used for" relationship, i.e., equivalent terms or synonyms
RT: "Related to" relationship, i.e., terms connected in the domain under consideration
SN: "Scope note", i.e., note on the use of a term or its definition

Table 1. Relationships in the thesaurus

The main thematic expansion algorithm (cf. Algorithm 1) uses two (parametrable) thresholds to decide the strategy to adopt.

Algorithm 1: Global algorithm for thematic expansion of the query

```
Data: the user query reqInitial, the thesaurus, threshold_low, threshold_high
reqModified;
if numResults(reqInitial) ≤ threshold_low then
    reqModified = expansionThema (reqInitial, thesaurus, "UF");
    if numResults(reqModified) ≤ threshold_low then
        reqModified = expansionThema (reqModified, thesaurus, "NT");
        if numResults(reqModified) ≤ threshold_low then
            reqModified = expansionThema (reqModified, thesaurus, "BT");
        end
    end
    expansionInteractive(reqModified, thesaurus);
end
else if numResults(reqInitial) ≥ threshold_high then
    reqModified = filteringThema (reqInitial, thesaurus, "NT");
end
execute (reqModified);
```

If the number of results obtained by the user's search is smaller than *threshold_low*, i.e., when we consider the number of results as insufficient, successive calls are made to an automatic expansion algorithm until a sufficient number of results are obtained (subject to a limit of 3 calls). This algorithm is called for a specific link type in the thesaurus (synonym, narrower term, broader term). The algorithm searches for all keywords of the initial query that are found in the thesaurus, and then looks for additional keywords in the thesaurus that match the specified relationship type with them. The keywords found are added to the query and it is re-executed. In case the number of results returned is greater than *threshold_high*, a filtering algorithm is called. This algorithm is interactive and, after consulting the thesaurus, offers keywords that are more specific than those of the user. The user can then choose or reject one or more of the proposed

terms and relaunch his query with the new keywords (Interface in Fig. 1).

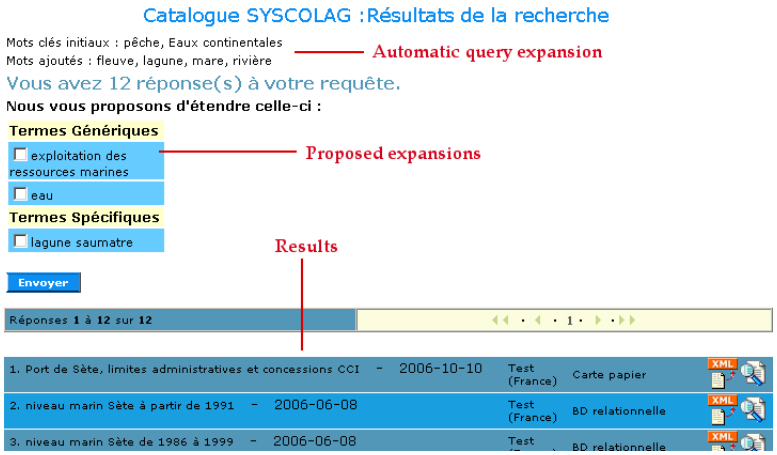

Fig. 1. Results interface with thematic expansion

For example, if the user provides *trawling* as search keyword and if the system does not return any result, the expansion algorithm will find that there exists a synonym relationship between the terms *trawling* and *trawl fishing*. A new query will then be executed with the keyword *trawl fishing*. If, on the other hand, the user selects a keyword that is too general such as *method of fishing*, which returns numerous results, the system can offer him terms to help narrow his query, such as *line fishing*, *net fishing*, *trawl fishing*, etc.

5.2 Spatial expansion

Another type of query expansion seemed interesting enough to us to implement in the context of georeferenced data. This expansion method relies on data's spatial aspect. Here it is a filtering algorithm that intervenes when the user searches with a known geographical location as a search criterion and is faced with too many responses.
A spatial search can be conducted by drawing a *minimum bounding rectangle* around the search zone on the cartographic interface, either by providing the rectangle's geographical coordinates or by choosing a geographical object (which also possesses a corresponding minimum bounding rectangle). In either case, the expansion method is based on this concept of minimum bounding rectangle and the use of topological relationships between rectangles.
In case of a search with spatial criterion, the obtained results are documents whose associated containing rectangle is in intersection with or touches that of the search zone. If too many results are obtained, the system proposes to

the user, based on topological relationships, to restrict the search to documents whose containing rectangle is strictly included in the search zone. Of course, it would be possible to use this algorithm with an other topological relationship, for example to expand the query.

5.3 Combined expansion: thematic and spatial

Finally, a combined expansion mechanism was developed by combining the use of the thematic reference base and of the spatial reference base. If the user enters a spatial concept as a keyword, i.e., a keyword attached to a geographic layer of the spatial database, then the expansion – or rather the filtering – can use the choice of a specific object on the geographic layer as a way of refining the search. For example, if the initial query includes a keyword such as the spatial concept *lagoon*, the system offers the user the opportunity of selecting one *lagoon* in particular from a displayed cartographical interface. This expansion is interactive; the user can very well choose to retain all initial responses.
Another improvement in the cataloguing service that seems interesting to us is to better exploit search results by using the representation of knowledge connected to these results.

6 Semantic presentation of results

Our idea is based on a *strong hypothesis:* the keywords present in the search results should allow the implicit knowledge to be extracted once a semantic network is established. This network will be constructed from the comparison of the keywords present in the result metadata records with the domain's ontological knowledge. In fact, it makes sense that the terms selected for annotating a record should represent concepts having, amongst themselves, a significant relationship. It is these links that we will try to retrieve from the structure of the domain's semantic representation (thesaurus, ontology, etc.) to be able to put together the semantic knowledge network existing between them. This network should allow domain specialists to verify the consistency of the expression of their knowledge (missing links or errors in link types in the thesaurus) or suggest semantically close words to the user so that he can refine his search.

Algorithm 2: Flooding initialization algorithm

```
Data: all the keywords extracted from the results
Result: the global variable H contains all the paths of the semantic
        network
for all m ∈ keywords do
    target ← keywords − m;
    create the PATH graph reduced to a single node m;
    flooding(m,target,PATH);
end
```

In MDweb, the result of a query is, as in most search engines, supplied in the

form of a list where each record fulfilling the criteria is summarized (title, date of record creation, catalogue where it is to be found, etc.). An initial task consisted of retrieving all the concepts (keywords) used to describe these records. For example, a query for "trawl fishing" gave all the keywords describing the result records: "marine fisheries", "trawling", "hakes". While this set of keywords constitutes a first summary of knowledge, the lack of a structure between them impairs interpretation. We therefore propose an algorithm that relies on the semantic reference base to find different "paths" between these words and thus to better extract the knowledge stored therein (see Fig. 2).

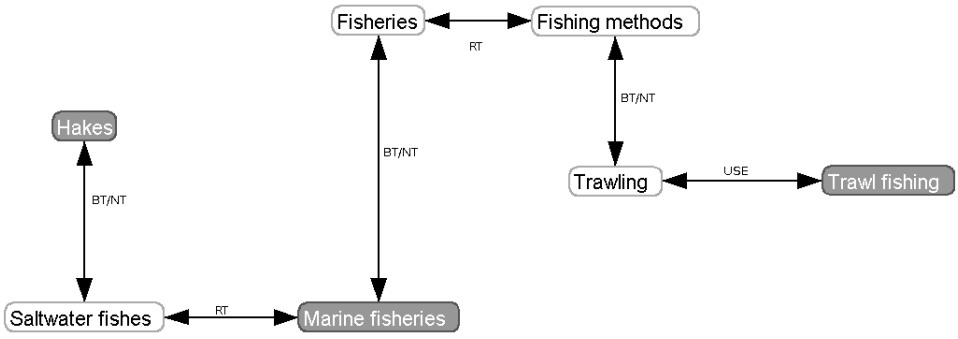

Fig. 2. Network of extracted knowledge (with the Prefuse visualization toolkit, http://prefuse.org/)

It is worth considering covering *all* the possible paths (especially if the reference base is of reasonable size) from different keywords present in the result to ascertain which amongst them arrives at another of these terms. The successive calls to the recursive algorithm 3 called by the algorithm 2 allows all the paths from one of these keywords (origin keyword) to be covered, and to store those that arrive at another of these keywords (destination keyword amongst the other keywords). We thus obtain the complete network with all the paths linking the keywords extracted from the result. Figure 3 shows one exploration stage of our example. Having started from the origin term "Marine fisheries", the algorithm arrives at "Trawling". On the figure, the greyed sub-graph represents the path already covered (the algorithm's PATH variable); the other nodes (in white) represent potential candidates for continuing the exploration (the successors of the "Trawling" node not yet processed). For each of these candidates, the associated node and the arc connecting it to PATH are added to PATH, and thus the exploration continues. When we process one of the search keywords ("Trawl fishing" in the example), the nodes and the arcs of PATH which are not already present are added to the global variable H, and the exploration stops (the paths originating from this other keyword will be processed in a different exploration.

Algorithm 3: Greedy flooding algorithm

Data: the current node c, the set of destination nodes d, the PATH graph of the path already covered since the first call

Result: at the end of recursive calls, the global variable H contains the graph made up of all the paths between node c and one of the nodes d, plus PATH

```
if c ∈ d then
    save PATH in H;
else
    for all nodes v successor of c in the thesaurus do
        if v does not belong to PATH then
            add the node v and arc (c,v) to PATH;
            flooding(v,d,PATH);
            delete arc (c,v) and node v from PATH;
        end
    end
end
```

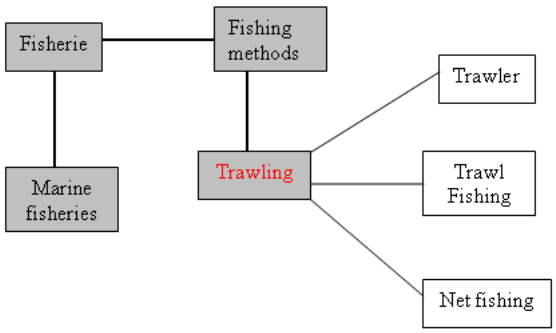

Fig. 3. Starting from "Marine fisheries", the algorithm processes "Trawling"

It soon becomes obvious that, due to its complexity, this solution is not viable for larger structures. A first method, already implemented, for optimizing the response time consists in limiting the path lengths by specifying a criterion of maximum path length covered, and thus abandoning terms too far away. An improvement over this method would be to find heuristics that would limit the exploration of the structure but would cover a sufficient number of significant semantic paths to construct a useful network. An idea is to follow only certain link types (for example, not to pursue SN (scope note) links of the thesaurus, see table 1), or to consider only certain link types either by selecting them in advance (for a specific search) or by relying on the immediate neighbourhood of

each term. However, these heuristics, albeit necessary for an acceptable response time, do return partially truncated information (we can see this on the example: if we do not follow RT links (see Table 1), the network of Figure 2 will not be found).

7 Conclusion

The use of cataloguing services in environmental applications is becoming indispensable. Nevertheless, these services will only have an impact on the communities concerned when they lead to semi-automatic input and, additionally, if search engines results become more relevant than they currently are. The propositions described in this article aim for this latter objective. The contribution of the semantic aspect as we have proposed it within the MDweb metadata service, by incorporating thematic and spatial reference bases to the ISO 19115 standard, is significant. Our ideas of query expansion as well as those of extracting knowledge implicit in the results could only be tested on a limited number of metadata records but show promise. The way ahead is to pursue the integration of the semantic aspect within the tool and the use of semantic co-constructions (enriching the thesaurus stored in a database by moving towards a shared ontology). For the results display, future improvments will consist to flag the different types of relationships (semantic, spatial and temporal). It should be used as a foundation for a new search module based on the refering thesaurus total display.

References

1. Barde, J.: Mutualisation de données et de connaissances pour la gestion intégrée des zones cotières. Application au projet Syscolag. Mastersthesis (2005) Université Montpellier II, Ecole Doctorale Information, Structures, Systèmes
2. Bottraud, J.C., Bisson, G., Bruandet, M.F.: Expansion de requêtes par apprentissage automatique dans un assistant pour la recherche d'information. CORIA (2004) 89–108
3. Bourigault, D., Lame, G.: Analyse distributionnelle et structuration de la terminologie, application à la construction d'une ontologie documentaire du Droit Traitement Automatique de la Langue **43-1** (2002)
4. Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology-Based Spatial Query Expansion in Information Retrieval. OTM Conferences **2** (2005) 1466–1482
5. Desconnets, J.C., Moyroud, N., Libourel, T.: Méthodologie de mise en place d'observatoires virtuels via les métadonnées. INFORSID actes du XXIeme congres (2003) 253-267
6. ISO23950:1998: Information and documentation – Information retrieval (Z39.50) Application service definition and protocol specifications, ISO 23950. International Organization for Standardization (ISO) (1998)
7. ISO19115:2003: Geographical Information Metadata, ISO 19115. International Organization for Standardization (ISO) (2003)
8. OGC: Catalog Service Specification. Open Geospatial Consortium Inc (2005)
9. Soualmia, L.F., Darmoni, S.J.: Projection de requêtes pour une recherche d'information intelligente sur le Web. RJCIA (2003) 59–72

10. Sowa, J.F.: Conceptual Graphs for a Data Base Interface. IBM Journal of Research and Development **20** (1976) 336–357
11. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American **284** (2001) 35–43