An open and powerful GIS data discovery engine

Authors

- **Martin Ouellet**, Library, Université Laval Québec, QC Canada, *Canada*
- **Stéfano Biondo**, Library, Université Laval Québec, QC Canada, *Canada*

KEYWORDS : semantic metadata mining, thesaurus, combined text-spatial search

The Geographic and Statistical Information Center (Centre GéoStat) of Laval University's Library is responsible for the acquisition, the organization and the dissemination of geospatial data. The center's objective is to support all students, teachers, and researchers across campus in their teaching and research work. The center manages more than 30 000 datasets, which represent about 10To of hard drive space. Like most academic libraries, its collection consists of raster and vector dataset useful for many different fields such as geography, political sciences, forestry, geology, health, soil science, marketing, history, topography, etc.

From a technological point of view, the "Big Data" phenomenon, of which geospatial data is part, will surely prove to be one of the landmarks of the 2010-2020 decade [2]. While data keeps growing, hardware gets cheaper and cheaper. Therefore, we have the means to store in-house entire collections of data. However, most organizations postpone the implementation of a discovery engine and leave unexploited this wealth of information.

In the past, geospatial data was associated only to a few specific disciplines, such as geography, engineering or land surveying. However, during the last decade, the use of geospatial data have spread because it became widely accessible for free (Open Data) and because more tools were developed to create, gather, exploit and disseminate this type of data (Google Earth, Open Source

software, « Mashup », iPhone and other GPS-enabled mobile technologies) [7]. The number of users grew exponentially and got more and more diverse (this can also be observed in academic settings). The statistics compiled by the Geographic and Statistical Information Centre show that today's clientele transcends disciplines. The new groups of users are generally inexperienced and are in need geospatial data for various and "untraditional" reasons: to back sociological or demographic studies, to plan itineraries, to help managers or company administrators take decisions, etc. In a nutshell, clientele is on the rise, clearly divided between novices and experts, and geospatial data can now be used for many different purposes.

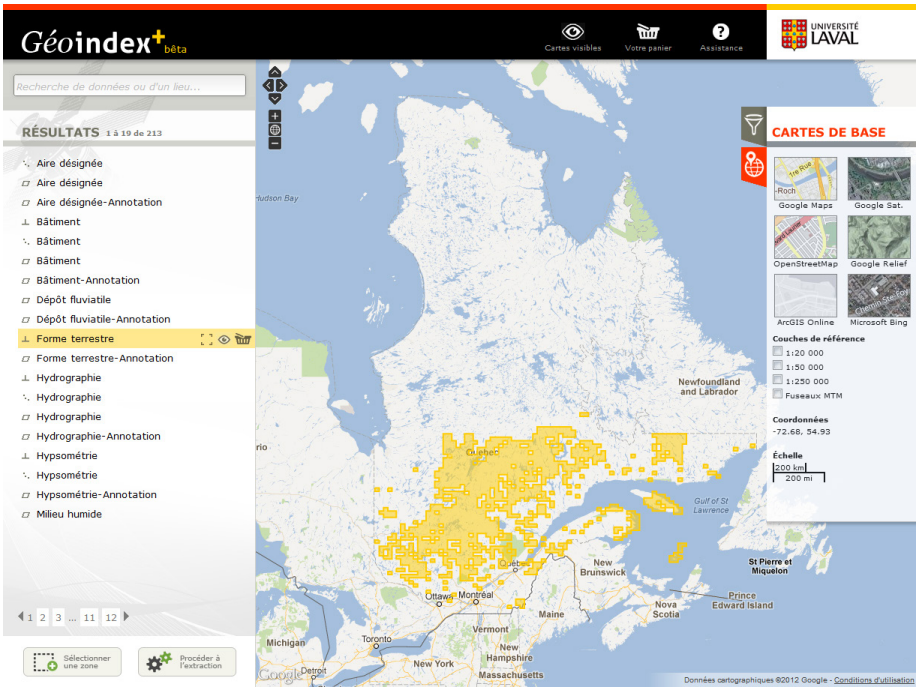


FIGURE 1

Screen capture of Geoindex+ interface

The challenge is to adapt to both realities: the ever-increasing amount of data and the changing nature of users. It is paramount to develop tools that are user friendly (to accommodate the various degrees of expertise) and powerful (to cope with the volume of data) to help users find what they need in this mass of information. Unfortunately, user-friendly search interfaces and better technologies don't always adequate with optimal outcome. Enabling semantic networks can significantly increase the relevance of search results, in accordance with users' needs. Therefore, the Centre GéoStat decided to develop a powerful discovery engine that combines traditional text-based search and spatial filter: Géoindex+ (Figure 1).

FIGURE 2

Screen capture of Geoindex+ simple search box

It features a simple search box (Figure 2) combining both metadata queries (to find relevant dataset) and the geocoding service API from Google (to help user locate on the map). If the user chooses to navigate directly on the map, the list of result will be automatically updated based on the current map extent.

The effectiveness of any search engine rests on the richness and quality of metadata. In March 2006, the Geographic and Statistical Information Centre set up a committee consisting of geospatial data users from Université Laval. The role of this committee was to enable and foster campus-wide access to geospatial data, to facilitate expertise sharing and to make the acquisition of most data possible. Work carried out by the committee led to the creation of a metadata framework that complies with the North American Profile of ISO 19115:2003 [1]. This allows for a standardized description of the various data sets, supports more effective data discovery and improves interoperability between different search engines.

We decided to use GeoNetwork, an geographic metadata catalog application that appears to be the most advanced open source product on the market [3]. It is an application with continuous development and a growing community. It is currently used in numerous Spatial Data Infrastructure initiatives across the world such as FAO, INSPIRE, WHO, etc. [5]. GeoNetwork has given us a web-based environment for editing and storing metadata within a database.

Special care was taken during metadata entry to ensure consistency in the choice of terminology. ISO 19115 allows for the selection of any recognized thesaurus [6]. We chose an encyclopedic thesaurus instead of a more specific one because our data set is multidisciplinary and not limited to a single academic subject matter. The Répertoire de vedettes-matière (RVM), a national standard for French-language indexing since 1974, met all of our needs [4]. Developed and maintained by Laval University, this encyclopedic thesaurus contains more than 273 000 entries, enhanced by semantic networks (specific terms, related terms, cross-references and English-language equivalents). It allows us to describe metadata using the appropriate vocabulary. In addition, the RVM is flexible and continually evolving by allowing us the possibility to modify or add new terms.

Inspired by the new generation of library discovery tools, we wanted to implement faceted searching in our engine to allow users refine results according to fields specific to GIS files. We generated indexes from our

GeoNetwork metadata records through the Apache Solr platform [9]. This powerful indexing technology has proven with large volumes and can easily support queries over millions of records. Considering that we will have a few thousand products listed in our system, the performance of the text search through our metadata is more than sufficient. This technology also made it possible to add to our discovery engine a "semantic autocompletion" search box and to enable full-text searching.

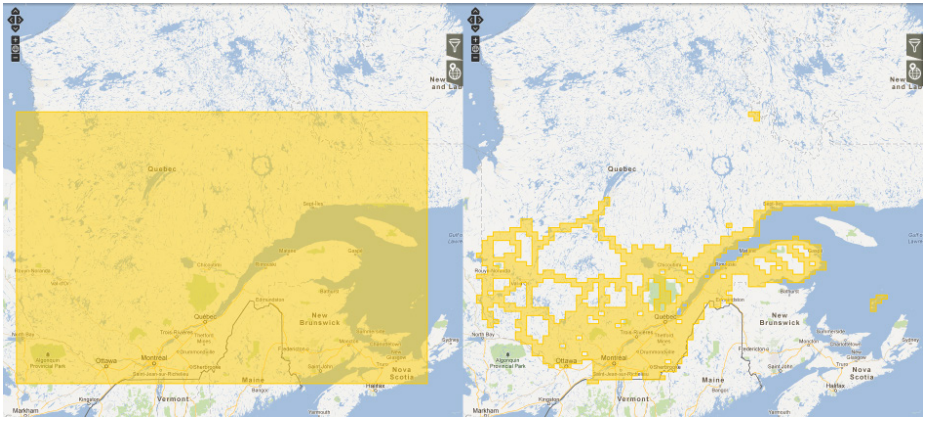


FIGURE 3

Default minimum rectangle versus more accurate complex polygon

ISO 19115 includes a section pertaining to the geographical scope of data sets. This scope is defined by the minimum bounding rectangle capturing the whole of the data set. Although essential to the formulation of a basic spatial search, the delineation of these boundaries can be considered too vague. In order to alleviate this problem, we created more precise and more complex delimitations for each data set, this to represent more faithfully the territory covered (Figure 3). These complex polygons were created using geometric operators from the PostGIS database [8]. Then we associated every metadata record to its geometric shape. It is therefore possible to refine a search by

moving across the map. We would like to keep this complex geometry directly in the metadata description instead of the default minimum rectangle because it is more relevant for the spatial filter. The XML implementation of the ISO19115 standard provides a definition for this type of complex polygon (in GML format), but it is not fully supported in the version of GeoNetwork that we use (2.6.3). To ensure that our solution is effective (at the end with the entire inventory that we have), we have simulated queries (with both spatial and textual filter) with nearly 3,000 metadata records combined with the same number of complex polygons. The response time was never more than a few seconds in every case.

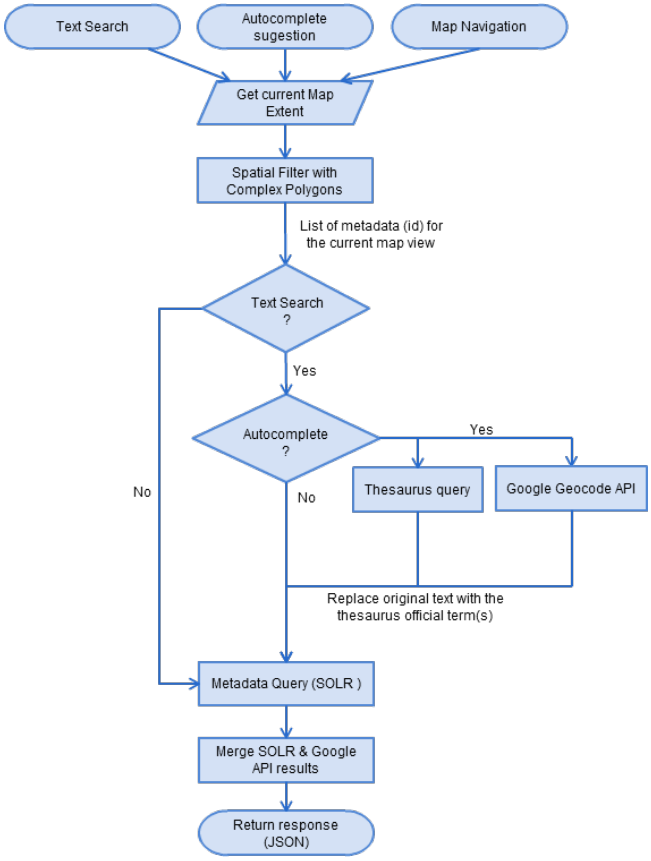
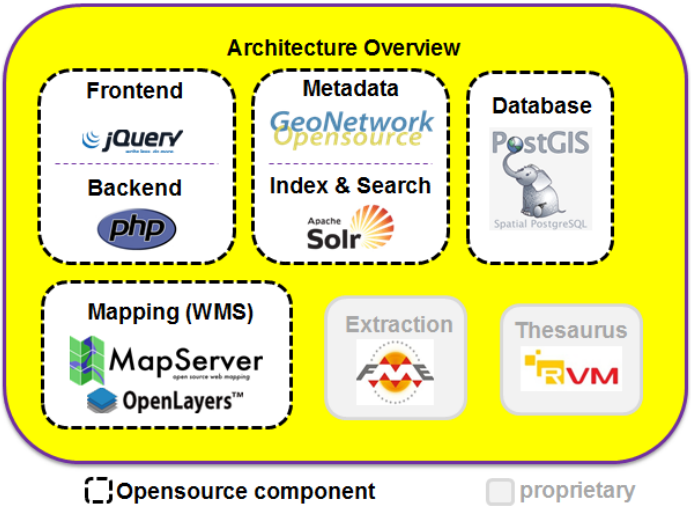


FIGURE 4
The search workflow

The search engine (Figure 4) we developed meets the challenge of helping various users to find what they need in a sea of data. It is also the corner stone of a bigger project: a web-based platform called Géoindex+ that is used to disseminate and extract geospatial data. The search capabilities and the geospatial data visualization section of this application is an assembly of existing opensource components (Figure 5). To put forward the characteristics of the search engine, demonstrations with Geoindex+ will be performed during the communication.

FIGURE 5

Architecture Overview



This communication is aimed to everyone who has to bridge the gap between a huge amount of geospatial information and ease of discovery. Those who are convinced of the importance of metadata to facilitate discovery, but that need to find time (and courage) to address this issue, will certainly benefit from this communication.

[1] CANADIAN GENERAL STANDARDS BOARD. North American Profile of ISO19115:2003 - Geographic information – Metadata. USA-Canada, 2008. Retrieved November 15, 2010 from http://www.cits.rncan.gc.ca/html/brodeurj/.protege/.napMetadata/review3/napMetadataProfileV1_2_1_en_final20081215.pdf

[2] DUMBILL, E. Planning for Big Data. O'Reilly Media, Inc.,2012.

[3] FEDERAL GEOGRAPHIC DATA COMMITTEE. ISO Metadata Editor Review, 2009. Retrieved December 2, 2009 from http://www.fgdc.gov/metadata/iso-metadata-editor-review

[4] GASCON, P. Le Répertoire de vedettes-matière de la Bibliothèque de l'Université Laval : sa genèse et son évolution. *Documentation et bibliothèques* 39, 3 (1993), 129-139.

[5] GEONETWORK OPENSOURCE. Gallery. Retrieved June 18, 2012, http://geonetwork-opensource.org/gallery/gallery.html

[6] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). ISO 19115 - Geographic information, metadata, Geneva, 2003.

[7] JONSHON, L., LEVINE, A. and SMITH, R. 2009. *The 2009 Horizon Report*. Retrieved December 18, 2009 from http://net.educause.edu/ir/library/pdf/CSD5612.pdf

[8] OBE, O. R. and HSU, LEO S. *PostGIS in action*. Manning, Stamford, 2011.

[9] SMILEY, D. and PUGH, E. Apache Solr 3 Enterprise Search Server. Packt Publishing, Birmingham, 2011.