Exploring a Geographical Dataset with GEOLIS

Olivier Bedel, Sébastien Ferré and Olivier Ridoux
IRISA/Université de Rennes 1
Campus de Beaulieu
35042 Rennes Cedex, France
firstname.lastname@irisa.fr

Erwan Quesseveur
RESO UMR CNRS ESO 6590
Université de Rennes 2, Campus Villejean
35 043 Rennes Cedex, France
firstname.lastname@uhb.fr

Abstract

Geographical data are mainly structured in layers of information. However, this model of organisation is not convenient for navigation inside a dataset, and so limits geographical data exploration to querying. We think information retrieval could be made easier in GIS by the introduction of a navigation based on geographical object properties. For this purpose, we propose a prototype, GEOLIS[1]*, which tightly combines querying and navigation in the search process of geographical data. GEOLIS relies on Logical Information Systems (LIS), which are based on Formal Concept Analysis (FCA) and logics. In this paper, we detail data organisation and navigation process in GEOLIS. We also present the results of an experimentation led on a real dataset.*

1. Introduction

For several years the amount of geographical data produced around the world keeps increasing. In the same time, a growing number of heterogenous geographical datasets became available online for watching or downloading through geo-spatial web portals . These portals enable to find specific layers of geographic information from a region of interest and a metadata description. Once you have got the desired layer, Geographical Information Systems (GIS) provide tools to explore and analyse it.

Geographical information is traditionnaly organised in two kinds of layers: raster layers and vector layers. Raster layers correspond to geo-localized images (e.g. satellite images, aerial photographies, digital elevation model) in which each pixel qualifies a portion of space with a unique value (e.g. infra-red radiation, elevation). Vector layers gather geographic objects, called features, which are qualified by a set a thematic attributes and spatially represented with geometric shapes (points, lines, and polygons) [6]. In the following, we will exclusively focus on information retrieval in vector layers, which by nature lend themselves to symbolic data analysis. Conversely, raster layers are traditionnaly processed using numeric methods which are not adressed here.

While GIS have recently known improvements concerning data storage (spatial databases, geo-spatial servers,...) and data display (webmapping services, globe visualization,...), the principles of information retrieval inside a geographic dataset keep unchanged. In fact, information retrieval in GIS is either querying like in database or browsing a set of layers.

Querying is powerful and expressive, but difficult to use when the goal is not very clear, and out of reach of many users. In GIS, thematic attributes are stored in tables attached to the layers. SQL-like languages are used to query these attributes, and spatial predicates enable to compute spatial relations between reference objects and searched objects, e.g. distance or inclusion. If thematic and spatial queries were used to be built using separated tools, the arrival of spatial function for SQL tends to unify querying interfaces.

Navigation in GIS is rigid and limited. As data are partitioned in predefined layers, users may check or uncheck layers to be explored and then graphically select on the map features or regions of interest. But then, no more navigation based on thematic attributes is possible. If browsing the table of attributes is possible, it does not help anyone in rapidly retrieving a specific information. All the more if the only available actions on the table consist in changing the columns order or sorting rows according to the values of a particular field.

To our point of view, querying, and more generally information retrieval in GIS could be made easier with the introduction of a real navigation among the attributes of geographical features. Our proposal is to assist the user with an incremental search process tightly combining querying and navigation. For this purpose, we have developped a

[1] This work is funded by a scholarship from Région Bretagne

prototype, called GEOLIS, which uses a Logical Information System (LIS) [4] that we have adapted to handle geographical data. LIS provide guidance in the search process by tightly combining querying and navigation in an incremental way. For data organisation and retrieval, LIS rely on Formal Concepts Analysis (FCA) and the expressiveness of logics.

FCA [5] organises data in a concept lattice. This lattice is a navigation structure that is automatically derived and adapts to changes in data. Concepts are subsets of objects sharing same properties. FCA properties are limited to boolean attributes whereas many documents, and especially geographical features, are described by valued attibutes. This is why in LIS, FCA has been generalized to Logical Concept Analysis (LCA), which allows to describe objects and concepts with logical formulas. This enables to deal with several domains of values in feature descriptions (e.g. integers, strings, coordinates, shapes) and to use patterns in queries (e.g. intervals, regular expressions, areas).

The sequel of the paper is structured as follows. Section 2 presents other uses of logics and FCA to handle geographical data. Section 3 details the organisation of geographical data used in GEOLIS. Section 4 describes the information retrieval process and briefly introduces GEOLIS architecture. Last, Section 5 presents the results of our prototype in the exploration of a real dataset dealing with the spatial distribution of rodents in Sahelo-Sudanian Africa.

2. FCA and Logics applied to GIS

FCA has not been yet really investigated for the management of geographical information. However [10] show the opportunity of combining conceptual structures and GIS for information access. Even if the final aim of this work is not to handle geographic space but abstract space such as the World Wide Web, it illustrates the representation of geographical information in a conceptual environnement. Another work, even if not directly connected to FCA, [7] illustrates the use of Galois lattices to represent and reason about topological relations on raster images.

Conversely, logical formalisms have been widely used to formalize spatial relations and especially topological relations [1, 3]. This also includes modal approaches to define the notion of proximity. However these approach are rather qualitative and often not directly useable with real world data. Description logics provide a framework more devoted to information retrieval, and have led to some encouraging attempts to describe the geographical domain. For instance, the VISCO system [11] relies on description logics to query a spatial database in a visual way. But if these systems integrate comparison, querying, and even query completion capabilities, conversely to LIS, they do not offer a relevant navigation structure derived from the data.

GEOLIS, and more generally LIS, combine a quantitative approach as data are described by expressive logics on concrete domains, and a qualitative approach, derived from the former, through conceptual structures and logical reasoning.

3. LCA to Organise Geographical Data

As mentioned previously (see Section 1), the layer structure is rigid. It imposes the same description schema for all the features of a layer. Moreover, as layers are considered as atomic objects in almost all GIS operations, working on a subset of geographical features often requires preprocessing and the use of temporary layers. Furthermore, using rows of an attribute table to store the description of geographical features is not convenient for navigation. A model with a thiner granularity would improve flexibility in geographical data handling. This is why with LCA, we propose to center the data model on the objects that are to be retrieved, i.e. the geographical features.

In the GEOLIS data model, each geographical feature is represented by an object and is described by a logical formula. This logical formula corresponds to the conjunction of logical properties derived from the attributes of the feature. A logical property is a valued property composed of the name of an attribute and of a corresponding value. For instance, here is a possible formula for describing an object o representing the French city Rennes:

```
d(o) = name:``Rennes''
AND population:206 000
AND geometry:POINT(351 869.83 6 789 643.91)
AND description:``administrative center
of the French region Brittany''
```

In this example, the conjunction operator `AND` combines logical properties defined on string values (`name` and `description`), on integer values (`population`), and on geometric representations (`geometry`). In GEOLIS, such a description is automatically extracted from layers encoded in common geographical data formats. For each feature of a layer, all attribute values are translated into logical properties. So is the geometry of the feature, using a standardized string representation, which enable to reason both on geographic shape and localisation. This way, the same language is used to query about spatial location and thematic attributes. This is a significant point to make information retrieval easier in GIS.

In fact, GEOLIS uses specialized logics (e.g. string logic, integer logic or spatial logic) to define the concrete domain associated with each attribute. Queries take the form of logical formulas, which may include logical properties whose values have been replaced by specific patterns. In addition to conjunction operator, queries may also include disjunction and negation operators OR and NOT. For instance, just consider the following query q:

```
q = population:>=100 000 AND
geometry:inside POLYGON(300 000 6 000 000,
400 000 6 000 000,400 000 7 000 000,
300 000 7 000 000, 300 000 6 000 000)
```

Spatial patterns are graphically drawn on a map, and then transcripted into a logical formula. In this example, the set of answers of q include the previously defined object o because the integer value 206 000 is more specific than the pattern >=100 000, and the spatial value POINT(...) is also included in the spatial pattern POLYGON(...).

Specialized logics define a partial ordering ($\sqsubseteq$) between values and patterns of a concrete domain, e.g. 206 000 $\sqsubseteq$ >=100 000. Logical properties may also be ordered themselves according to a taxonomy specific to the dataset explored. For instance, the rodent dataset which is fully introduced in Section 5, gives information about biometry, location, phylogeny and date of capture of rodents that have been trapped and observed in the Sahelo-Sudanian stripe. In this dataset (see Figure 1), we have the following relations: $sex \sqsubseteq biometry$ and $age \sqsubseteq biometry$.

4. GEOLIS to Explore Geographical Data

In GEOLIS, LCA enables navigation through all properties of the geographical features. In LCA, a concept groups together objects sharing a common set of properties, the intent, and can be seen as a query. An edge of the concept lattice is a navigation link from a concept to another, i.e. from a query to another query. In other words, an edge corresponds to a property which is a query increment. In fact, querying consists in being on a particular concept, and navigating consists in reaching a neighbour concept by following an edge of the lattice. However, in LIS, the complete concept lattice is never built. At each step in the search process, the LIS kernel computes only the increments relevant to the current concept, i.e. to the current query. By analogy with the working directory of file systems, we call the current query the *working query* (*wq*). Formally, if $ext(wq)$ denotes the features satisfaying wq, GEOLIS will provide all query increments p belonging to: $max_{\sqsubseteq}\{x|\emptyset \subset ext(wq \texttt{ AND } x) \subset ext(wq)\}$. Following query increments reduces the number of answers, but never leads to dead-ends. Furthermore, only the most general refining increments are provided. As a large amount of query increments may be confusing for navigation, GEOLIS organises them in a navigation tree, according to the taxonomy on properties and to the partial order between values and patterns of properties. The tree is progressively built during the navigation. At the beginning, only the first level is computed.

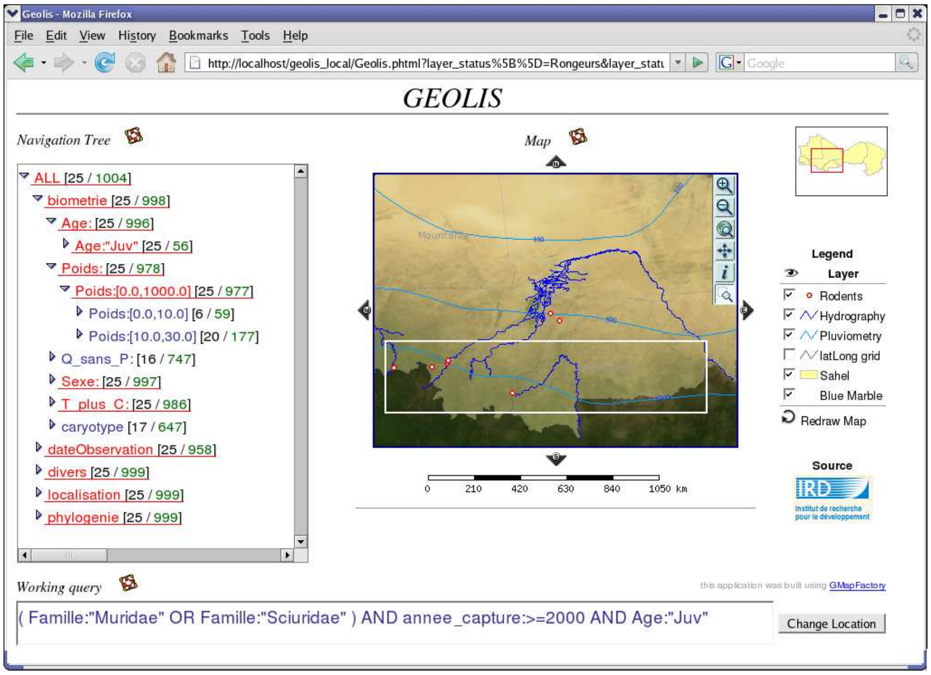
4.1. GEOLIS Interface

The GEOLIS interface, which is shown in Figure 1, is composed of three main parts: the *navigation tree* placed on the left, the *map area* filling the center and the right, and the *working query box* at the bottom.

- The *working query box* displays the current query in the navigation. It indicates the query describing objects rendered in the map. The query box is editable, so that it is possible to enter manually a new query or to modify the current one.

- The *map area* is a composed component. A main map including fixed background layers (e.g. administrative boundaries, hydrography or satellite image) indicates by white points the location of geographical features satisfying the current query. A legend details symbology of the main map and enables to specify which layers are visible. A keymap locates the boundaries of the main map on a world map. Last, standard map tools are also available: pan, zoom in/out and *to full extent*. The *map area* component comes almost unchanged from an existing interface. It has been enhanced with a *logical zoom* tool, which enables to select rodents directly on the map by drawing a rectangle, i.e. a bounding box, enclosing them. The *logical zoom* produces graphic query increments that are expressed in queries with the syntax geometry:inside POLYGON(...).

- The *navigation tree* is a visual representation of the partially ordered set of query increments. Query increments are properties shared by at least one feature of *wq*. Each node of the tree represents a query increment which can be used to change wq (see Figure 1). When a node is expanded, this entails the computation of query increments that refine the increment of this node. Then, the new increments appear as its children in the tree. The root of the tree is ALL, i.e., the most general formula. Nodes under the root correspond to general properties of the taxonomy built over the dataset. Then nodes represent pattern properties or value properties which are the leaves of the tree. Each node of the tree is rendered with an icon, a label, and two numbers. The label is the formula representing the increment. The style of the label is also informative. Underlined orange labels correspond to formulas shared by all the rodents of wq, whereas blue labels indicate properties that discriminate them. The two numbers indicate a proportion: the count of rodents in wq that the increment leads to, i.e. the support, out of the count total of rodents sharing the formula. Two actions are possible in the tree: (1) collapsing or expanding a node by acting on the icon, (2) updating wq by selecting a label.


Geolis - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost/geolis_local/Geolis.phtml?layer_status%5B%5D=Rongeurs&layer_statu
Google
GEOLIS
Navigation Tree
Map
Mauritania
0 210 420 630 840 1050 km
Legend
Layer
Source
IRD
Institut de recherche pour le développement
this application was built using GMapFactory
Working query
Change Location

- ALL [25 / 1004]
  - biometrie [25 / 998]
    - Age: [25 / 996]
      - Age:"Juv" [25 / 56]
    - Poids: [25 / 978]
      - Poids:[0.0,1000.0] [25 / 977]
        - Poids:[0.0,10.0] [6 / 59]
        - Poids:[10.0,30.0] [20 / 177]
    - Q_sans_P: [16 / 747]
    - Sexe: [25 / 997]
    - T_plus_C: [25 / 986]
    - caryotype [17 / 647]
  - dateObservation [25 / 958]
  - divers [25 / 999]
  - localisation [25 / 999]
  - phylogenie [25 / 999]

☑ Rodents
☑ Hydrography
☑ Pluviometry
☐ latLong grid
☑ Sahel
☑ Blue Marble
Redraw Map

(Famille:"Muridae" OR Famille:"Sciuridae") AND annee_capture:>=2000 AND Age:"Juv"

Figure 1. The GEOLIS interface. Selected rodents appear as white points on the map. The white rectangle on the map represents the region that has just been selected as a graphical query increment.

During the navigation process, the interface is always maintained coherent. Each action on the *working query box*, the *navigation tree*, or the *map area* entails the update of all the components. Figure 1 illustrates the result of this interaction during the exploration of the rodents dataset. The query focuses on the rodents of families "Muridae" and "Sciuridae", trapped since 2000. So, by editing the working query box and using the `OR` operator, we restrict navigation to families of interest. In the same way, we limit investigation to rodents captured since 2000, using the pattern `annee_capture:>=2000`. Then, in the navigation tree, we select the increment `Age:``Juv''` to keep only young rodents. Notice that we could have the same result by manually editing the working query box anew. Figure 1 shows the state of the interface at this stage of navigation. The map and the navigation tree have been updated, w.r.t. the current query. We have the possibility to reduce the number of rodents by selecting a weight range in the tree, but we choose to focus on the set of rodents in southern Malia. With the *logical zoom* tool, we draw on the map a rectangular shape enclosing the desired region (see Figure 1). The logical zoom is not a graphical zoom. It will not modify the extents of the map, but it will entail the update of the current query, and consequently of the navigation tree and of the features drawn on the map. The rectangular shape will be translated into a formula based on the `geometry` property, which will be automatically added to the current query:

```
wq =(Famille:"Muridae"
OR Famille:"Sciuridae") AND
annee_capture:>=2000 AND Age:"Juv" AND
geometry:inside POLYGON(-14.056 10.0585,
0.540 10.0585, 0.540 13.393,
-14.056 13.393, -14.056 10.0585)
```

The navigation tree will be reduced, and show only properties and increments concerning rodents of the selected area. Last rodents in the north of the Burkina Faso border will disappear from the map.

4.2. GEOLIS Implementation

The GEOLIS prototype results from the coupling of several technologies from LISFS [9], web mapping and web domains.

LISFS is a generic implementation of LIS, and is at the same time a genuine Linux file system [9]. In LISFS, files and file parts (lines) are objects, paths are queries, directories are navigation places, and subdirectories are the automatically computed query increments. Two kinds of plugins can be used in LISFS: logics and transducers. Logics define the kind of formulas that can be used in object description, and queries. Transducers allow to partially automate the description of objects, depending on the file format. The Geographical Markup Language (GML) proposed by the OpenGeospatial Consortium is the data format we chose to use in GEOLIS. For our purpose, it has the advantages to gather all information in one file whose XML based structure may be rearranged w.r.t. to GML specifications. The GML transducer automatically extracts the spatial description and the thematic properties for each geographical feature o stored in the GML file, and produces a description formula $d(o)$ (as illustrated in Section 3).

LISFS constitutes the kernel of GEOLIS, where the geographical data to be explored is stored. The GEOLIS graphical interface is a web interface. The navigation tree and the working query box have been designed using the server side language PHP. The map area is built with the widely used map generator UMN MapServer, which renders geographical features stored in GML files. Interaction between GEOLIS components is detailled in [2].

5. Experimentations

GEOLIS have been tested out with a real dataset qualifying the distribution of rodents in the Sahelo-Sudanian stripe. This dataset results from the merging of several databases provided and maintained by the French Institute for Research and Development (IRD[2]) since 1980 [8]. It aims at determining possible causes affecting the distribution of rodents, but as data comes from local observations, this base is an imperfect sampling of Sahelo-Sudanian stripe. It gathers more than 20 000 features, each one being described with properties about biometry, philogeny, period of capture and localisation.

First experiments highlighted several occurences of anomalous entries resulting from errors in data collecting and merging. In the navigation tree, all values of a properties are listed with their count. This way, we can directly determine the most likely domain of a property ({``F'', ``M''} for Sexe:), and identify synonymous (``m''), uncertain values (``M?'') or mistakes (``49''), in order to correct them.

The unperfectness of the sampling has also been observed during the navigation. For instance, just by looking at the count of rodents under each value of the node pays (i.e. country) in the navigation tree, we noticed that half of information in the base comes from Senegal, while it represents only a small part of the studied area on the map. Having knowledge about data origin could enable to balance future results concerning rodent distribution. So we decided, as a first step, to use GEOLIS to qualify underlying strategies in the sampling of the database. For instance, within the navigation tree, we can restrict navigation to rodents trapped alive, and visualize, at the same time, properties annee_capture (i.e. year of capture) and habitat. In only a few operations, we noticed that for rodents captured alive, the diversity of trapping places tightly depends on the period of capture. For instance, 85% of rodents found in savanna were trapped in year 2000, which also corresponds to more than 60% of rodents trapped that year. Less diversity in trapping places could be explained by trapping sessions led to confirm hypotheses implying location criteria.

6. Conclusion

With the expressiveness of specialized logics and LIS exploration paradigm, GEOLIS enables a tight combination of querying and navigation applied both on thematic and spatial properties of geographical data. Not only information retrieval is improved, but navigation tools also enable to discover underlying properties of the explored dataset. Furthermore, experiments have been successfully conducted with real data expressed in a common GIS data format. In the future, we plan to integrate spatial relations in GEOLIS to improve expressiveness and querying capabilities.

References

[1] N. Asher and L. Vieu. Toward a geometry for common sense: A semantics and a complete axiomatization for mereotopology. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995.

[2] O. Bedel, S. Ferré, O. Ridoux, and E. Quesseveur. GEOLIS: A logical information system for geographical data. In *Int. Conf. Spatial Analysis and GEOmatics - SAGEO*, 2006.

[3] A. G. Cohn. Qualitative spatial representation and reasoning techniques. In *KI '97: Proceedings of the 21st Annual German Conference on Artificial Intelligence*, 1997.

[4] S. Ferré and O. Ridoux. An introduction to logical information systems. *Information Processing & Management*, 2004.

[5] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, 1997.

[6] R. Laurini and D. Thompson. *Fundamentals of Spatial Information Systems*. Academic Press Limited, 1992.

[7] F. Le Ber and A. Napoli. Design and comparison of lattices of topological relations based on galois lattice theory. In *Eight International Conference on Principles of Knowledge Representation and Reasoning - KR'2002*, 2002.

[8] L.Granjon. Inventaire et caractérisation des espèces de rongeurs sahélo-soudaniens. Technical report, IRD, 2007. http://www.mali.ird.fr/activites/inventaire.htm.

[9] Y. Padioleau and O. Ridoux. A logic file system. In *Usenix Annual Technical Conference*, 2003.

[10] U. Priss and L. J. Old. Information Access through Conceptual Structures and GIS. In *American Society for Information Science Conference (ASIS'98)*, 1998.

[11] M. Wessel, V. Haarslev, and R. Möller. Visual spatial query languages: A semantics using description logic. In *Diagrammatic Representation and Reasoning*. Springer, 2000.

[2] The authors would like to thank M. Laurent Granjon and M. Jean Marc Duplantier from IRD (CBGP UR 22 Montpellier) for their active contribution in the building of the rodents database.