Geographical information access for non-structured data

Julien Lesbegueries
LIUPPA
Université de Pau et des Pays de l'Adour
64013 Pau Université Cedex
julien.lesbegueries@univ-pau.fr

Mauro Gaio
LIUPPA
Université de Pau et des Pays de l'Adour
64013 Pau Université Cedex
mauro.gaio@univ-pau.fr

Pierre Loustau
LIUPPA
Université de Pau et des Pays de l'Adour
64013 Pau Université Cedex
pierre.loustau@univ-pau.fr

ABSTRACT

This paper presents the Virtual Itineraries in Pyrenees (PIV) project. Spatial and temporal unified models are proposed to give a formal representation to geographical information. The aim is to improve the access to local cultural and heritage document collections. The models take into account characteristics of heterogeneous human expression modes: written language and captures of drawings, maps, pictures, etc. Semantic treatments have been built to automatically manage spatial and temporal information from non-structured data. These treatments are added to classical information extraction (IE) approaches. Then, geographical information retrieval processing is based on geographical information systems (GIS) algorithms. These algorithms look for any relations between formal representations of geographic information in documents collections and similar representations in a user query. Finally we propose a prototype implementing such geographic IE and geographic Information Retrieval (IR).

Keywords

unified geographic feature model, non-structured data, semantic processing, content-based information access, heterogeneous expression modes

1. INTRODUCTION

The Virtual Itineraries in Pyrenees (PIV) project purpose consists in managing a repository of electronic versions of books, newspapers, postal cards, lithographs of the XIX^{th} and XX^{th} Century. These corpora are yet quite unknown and are only accessible in local area archives of museums and libraries. We want a non-expert user (teacher, learner, tourist or scholar) to better access these corpora. It is the reason why a regional media library supports this project. These cultural and heritage document collections are characterized by contents strongly attached to local areas and their land history [7].

The aim of our work is to make a content retrieval process more efficient each time a query includes geographical restrictions. Exploiting such topics will produce a higher relevance score in the PIV system.

Generally, spatial information is either supported by Relational Data Base Management Systems (RDBMS) and Geographical Information Systems (GIS) for structured data management or, by Electronic Document Management Systems (EDMS) and Library Management Systems (LMS) for semi-structured and non-structured data. All these systems aim to provide fast and effective content-based access to a large amount of information. But unlike GIS or some RDBMS software, EDMS or LMS software do not offer high-level spatial operators and use classical information extraction (IE) and information retrieval (IR) approaches generally statistical for their indexing method and their query language. Accordingly they are not sufficient to manage information in which semantics depends on spatial concepts. Thereby the results of queries about spatial information are generally disappointing.

To overcome these limits we propose a semantic approach to analyse and interpret spatial and temporal information contained in documents (or in queries). It's a way to manage geographical information more accurately [17, 23, 18]. So, PIV system integrates basic services of existing LMS and/or EDMS and new services marking and retrieving spatial and temporal aspect of information. It uses a specific architecture based on web services, spatial and temporal unified models to represent Geographic Features (GFs) and XML indexes to better manage geographic marks. The originality of our approach consists in spatial and temporal unified models. That allows to formalize every geographical information whatever its expression mode is (*i.e.* texts, maps, images). Moreover we propose a recursive strategy of complex GFs representation.

Geographical contents management within heterogeneous documents collections is the main purpose of the paper. We first focus on the geographical information semantic process in our repository of documents. Then we argue how PIV system improves information retrieval accuracy each time a query contains spatial criteria.

2. SEMANTIC PROCESSING

Geographical information in a document repository like the PIV one is distributed across various expression modes such as text, maps, tables. Each mode have specificities

SAC'06 April 23-27, 2006, Dijon, France

regarding the kind of geographical information they have to express. A text is more effective to explain facts in relation with a geographic place (as a named entity) than it is to describe the complex spatial organization of a phenomenon. In this case a map is much more efficient. However, the notion of time and evolution, difficult to render on an image or a static map, is naturally conveyed by text or graphics, such as curves better suited for showing the evolution of a phenomenon.

In such corpora, a GF is composed of a Spatial Feature (SF), a Temporal Feature (TF) and a phenomenon (Figure 1). The example "churches of the XVth Century at 8 miles in the south of Pau" is a perfect example of a complete GF (Figure 1). Let us assume that to set of a geographical retrieval process of such corpora SF have to be explicit. TF could be implicit or not localy expressed or may have range on more than one SF. Phenomenon can be missing. Consequently to process geographical information in-depth analysis of spatial information is mandatory. We rely here on a semantic processing approach that has been developed for several years and proves signicant results ([22, 9, 28]).

2.1 The "target/site" concept

Linguists' works explain human particular manner of representing spatial information in written language. According to [6], we can link a place to a category and associate it to a natural or artificial boundary. Four categories relative to our corpus can be specified: named boundaries (countries, counties, parishes ...), hydrographic features (rivers, estuaries, lakes ...), man-made features (cities, towns, villages ...), and physiographic features (mountains, plains, coast line ...).

Refering to such places involves several elements. [26] studied this assumption in written langage and propose the *target/site* concept. In written language the target and the site corresponding to a spatial evocation respect a position in sentence. When the target corresponds to the subject, the site corresponds to the object. More generally, in his hypothesis the target corresponds to the subject of our description and the site corresponds to its spatial and temporal references.

Our assumption is to extend this hypothesis to any other expression modes.

In this context, we focus here on the notion of recursive spatial and temporal absolute or relative location, a bounded but impossible to circumvent type of information carried by geographical information. In other words we focus on discursive or graphical structures combining several geographic entities, called here Geographic Feature (GF).

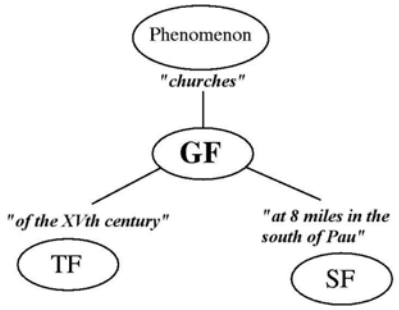

Figure 1: The composition of a GF.

```
Dans la première moitié du XIXe siècle, lors d'importantes
neiges dans les montagnes du sud-ouest de la France,
dans quelques villages basques des Pyrénées-Atlantiques...
(1) dans les montagnes du sud-ouest de la France
(2) quelques villages basques des Pyrénées-Atlantiques
(3) Dans la première moitié du XIXe siècle
```

Figure 2: Example of text expression mode.

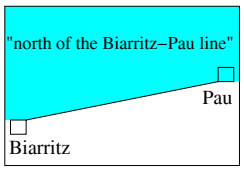

Figure 3: The Biarritz-Pau line in a graphical mode.

We claim that automatic analysis of such structures in our document repository provide an interesting indexation mode for querying such documents.

2.2 Unified Models

The semantics of absolute or relative locations (represented as feature structures) are shown in Figure 6.

In Figure 2, (1) expresses an exhaustive determination selecting all spatial entities of the given type mountains ("montagnes") located in a given zone, which matches the south-east half of the named geographic entity (France). In (2) the determination (introduced by "quelques" (some)) is relative, *i.e.* only a part of the elements given by the type has to be considered. Here, the type specifies that we only keep "Basque" villages from a given zone ("Pyrénées Atlantiques").

In (3) only a part of the XIXth element of the temporal type Century ("siècle") has to be considered. The part of the element is given by first part ("première moitié").

In fact in our corpus geographic expressions could be significantly more complex (as geographic feature defined by geometric zone: "north of the Biarritz-Pau line." Figure 3). Therefore in order to build an efficient geographical information retrieval (GIR) process [24, 29], specific spatial and temporal models are proposed.

These models have been thought to be compliant with geographical information contained in our repository. This information is represented in a non-structured form, polysemic and sometime context-dependent. Thereby the proposed semantic process for analyzing GFs gives a less formal representation than those needed in the world of GIS but enough for the next stage of the system, the information retrieval process.

Spatial axis

Contrary to [16], [15], [19] or GML[1] that manage well-formed spatial features (from the databases' point of view) we have to manage spatial features (SFs) expressed in different modes (written language, maps, images, etc.). Therefore, we de-

[1]Geography Markup Language - http://opengis.net/gml - However we use in our model a GML-based language to describe the geo-location of SFs.

fined a unified model to formally represent complex SFs containing unstructured spatial information. In our model, according to the linguistic hypothesis, the SF part of a GF is recursively defined from one or several other SFs and topological relations are part of the definition (figure 4). The [26]'s ideas can easily be defined in a recursive way. For instance, the SF in a literal expression like "north of the Biarritz-Pau line" :

- is first defined by sites (here 2 named entities) "Biarritz" and "Pau" that are well known locations,
- then the term "line" expresses a linear relation between the two sites cutting the landscape into two parts ("the Biarritz-Pau line"),
- finally an orientation relation determines the sub-space to focus on.

The temporal feature is implicit here. It can be identified by the document time or by a temporal feature introduced at the beginning of the paragraph, in the chapter title, etc.

This idea remains valid for other expression modes. On the one hand, named entities, as "Biarritz" or "Pau", can be associated to ponctual symbols : an icon for instance. On the other hand, complex SF could be expressed by using other graphics features : two contrasted areas to evoke a partitioning for instance as shown in figure 3.

So a SF can be (Figure 4):

- an Absolute Spatial Feature (A_SF) if it only consists in a named entity allowing a geo-localization,
- or a Relative Spatial Feature (R_SF) if it is defined using a topological relation with at least one SF. Topological relations can be adjacency, inclusion, distance, geometric and orientation [12], [13]. For instance :
  - an adjacency relation appears when we evoke a SF by spatial proximity with another SF. This relation is evoked in written language with terms like, near, close by, step by step..., as "near Laruns village" where the whole expression is a R_SF whereas "Laruns village" is an A_SF.
  - a geometric relation appears when we need to evoke several SFs to define a spatial feature by the evocation of a geometrical figure: *i.e.* "The Biarritz-Pau line" or figure 3.

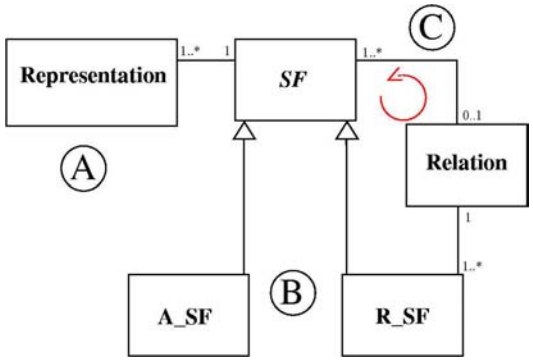

Figure 4: Unified Spatial Model simplified schema.

All these relations have attributes in order to characterize them. For example a relation of distance has a numerical parameter, a relation of adjacency has a qualifier.

All the resulting SFs are conform with the PIV unified spatial model. Thus, we can manage A_SFs like "in Laruns" and relative ones like "near Laruns", "at about 10 km in the south of Pau city", "between Pau and Laruns", etc. If we take back example "north of the Biarritz-Pau line": it is GF composed of a R_SF defined by the "north of" relation and the "Biarritz-Pau line" R_SF. This second R_SF is defined by the relation "line" and by the GF "Biarritz" and "Pau". So SFs extracted from various expression modes can be formally represented thanks to the unified spatial model.

Although this model has been built thanks to linguistic ideas on spatial reasoning, a similar schema can be modelled for time reasoning [4].

Temporal axis

As said in [10], the TFs could be implicit or not localy expressed or may have range on more than one GF. Anyway in our corpus when a TF is expressed it is an expression relating to a historical time. The expressions studied are either temporal periods (as "in the first part of the XIXth Century") or durations (as "for 80 years"). Note that our work in progress has not yet approached the question linked to the temporality of the events. Because of its importance in queries only expressions evoking the anchor in a chronology have been considered. For these reasons a very traditional method in formal semantics [8] seems to give sufficient results for our spot. The first step translates the temporal expression in a formal structure representing three types:

- algebraic intervals with a hierachical relation as Allen's [1] interval, such structure cleary appears in the expression "the first part of the XIXth Century";
- metric as in the expression "for 80 years";
- operation of succession as "in the next period".

The second step implies a contextual and referential interpretation to produce a "unified" interval between two dates.

This approach is similar to the spatial one so that a similar recursive model can be built. The temporal model is composed of a TF that can be an A_TF (a date) or a R_TF. A R_TF is composed of at least one A_TF and of some imbricated relations. For instance the R_TF "between 1914 and 1918":

- is first defined by two A_TF "1914" and "1918",
- then an interval relation "between" links these A_TF to focus on the closed interval.

Temporal relations can be closed or open interval relations ("between", "during", etc.) or distance relations ("before", "after", "10 years ago").

The particularity of the temporal axis is that an A_TF can be hard to define. Indeed it can be an implicitly inferred date. So a complex multi-granularity processing must be envisaged in order to find A_TFs introduced at the beginning of paragraphs or chapters.

2.3 The analysers

Let us retain that as the words or expresions in the written language, in images, maps or graphics it is obvious that the important semantic information necessary to interpret them is not represented in single pixels but in meaningful image objects and their mutual relations. These meaningful objets will be named from now on "sems". [25] [28] proposes semantics definition to represent spatial data. Finally *eCognition system* provides a powerful toolkit for image analysis[2].

An interesting characteristic of maps is the fact that they follow a rather strict structural pattern, since their construction by humans is more or less guided by formal rules, or at least by identified usages. Therefore, the use of the semiotic approach of information representation as studied in [3] allows to derive model of the map that will serve as a basis for the analyser tasks. Generally a document content processing sequence is composed of 4 main steps :

1. the "tokenisation" carries out a segmentation of the document in smallest sems;
2. the lexical and morphological analysis proceeds to a sem recognition;
3. the syntactic analysis, based on grammars, allows to find the bonds between sems;
4. finally, the "semantic" step carries out a more specific analysis allowing meaningful sems groupements to be interpreted.

In our data processing approach steps 3 and 4 are quite different. According to [2] we adopt an active reading behaviour, that is to say sought-after information is *a priori* known. Thus a pattern analyze is first performed to fetch "kernel" sems of the semantic expression to mark, thanks to a definite clause grammar (DCG) implemented in Prolog, both syntactic and semantic analyses are then performed. So AGFs (*i.e.* well known sites) are extracted first. Then RGFs are built from pointed out AGFs.

Prolog proves to be an interesting choice here since it allows unication on feature structures as well as other complex semantic computations to be integrated in the grammar, thanks to GULP [14].

2.4 Multi-indexation

As our system architecture is open (weak composed), based on web services, we can easily integrate specific tools according to needs. Indeed, an indexation layer is built for each semantic axis. So we can define several models and treatments and develop a dynamic multi-indexing system: a model definition and a specific grammar are enough to automatically build a new indexation layer.

More precisely, when some documents are added, they are marked and there is an identifier for each object, namely paragraph, part of an image or specific layer in a map. Then each extracted feature (corresponding to a semantic aspect) is depicted with its description and an object identifier. We can thus retrieve it by a pattern matching algorithm based on the description and applied on the object pointed by the identifier. A new semantic aspect management (temporal, educative, discursive, etc.) will generate a new file containing extracted features (corresponding to a pre-defined model) located in document by object identifiers.

We can see for example in section 3.2.2 a file extract created for spatial semantic aspect. It must validate the unified spatial model implemented in a XML Schema.

[2]http://www.pcigeomatics.com/products/definiens.html

```
root(X) --> GF(X).

GF(geographic_feature:X)-->gf(X).

gf(rgf:X) --> rgf(X).
gf(agf:X) --> agf(X).

rgf(relation:X..rgf:RGF) --> relation(X), rgf(RGF).
rgf(relation:X..agf:AGF) --> relation(X), agf(AGF).

relation(adjacency:X) --> adjacency(X).

adjacency(type_adj:close)-->ls_token('near').
adjacency(type_adj:close)-->ls_token('close').
adjacency(type_adj:close)-->ls_token('to').

agf(X) --> prep,lexicon(X).

prep --> ls_token('in').
prep --> ls_token('of').

lexicon(label:N..type:village) --> N@gf:agf..cgf:yes.
```

Figure 5: Spatial grammar extract.

3. PIV SYSTEM INFORMATION EXTRACTION AND RETRIEVING PROCESSES

PIV system implements IE and IR complementary approaches to better manage a cultural and heritage textual corpus. We need to search into a collection of documents (non-structured data for spatial computation usage), GFs semantically related to other GFs detected in a free text query. Then, it will be necessary to extract fragments of these documents, to classify them and, finally, to present them to the user.

As previously evoked our "weak composed" system based on web services easily integrate specific tools according to needs. Thus, our system manage geographic data for spatial semantic with GIS, via developped web services. Actually, GIS web services tools are used first to validate GFs candidates in the semantic processing stage. Then we use them to compute the GFs' geometric forms and geo-localization during the indexes creation stage.

3.1 PIV system information extraction

At the moment in PIV system for the written language all the IE process for SFs is fully implemented thanks to the Linguastream platform[3] [5] [27]. The writing of grammar for the temporal expressions is in hand via the same platform. The image processing is not yet implemented but thanks to the unified geographic model, a partial manual indexation is enough to take into accounts image documents. For now only a little corpus has been indexed to test PIV project: around ten documents and fifty lithographs containing one hundred or so extracted features. Concerning the "phenomenon" our idea is not for the moment to define a semantic IE process but to use the existing meta-data stored for each document thanks to a LMS. Our system allows to integrate results given by such a LMS in order to combine them with GF indexes.

[3]http://www.linguastream.org

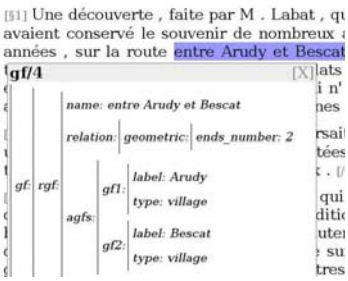

Figure 6: Extracted SF and its line semantic structure.

From now on only the SF process will be detailed. Document contents are taken into account by a specific semantic process that focuses on spatial features.

The data processing sequence used to detect spatial features is described in [23]:

The Figure 6 shows an example of semantic feature extraction result, obtained thanks to Linguastream platform. The extracted SF "entre Arudy et Bescat" *(between Arudy and Bescat)* is interpreted as a R_SF, defined by a geometric relation and 2 A_SFs "Arudy" and "Bescat" that are french villages.

There is a gap between databases data structures and semantic feature extraction for information management - more precisely between GISs data structures and our spatial semantic feature extraction system. Our system architecture allows lacks of information in SFs' definition and can manage incomplete ones. It can call additional services in order to complete these lacks.

Using the XML technology, we build index files from this marking tool (see next section). GIS tools provide a solution for candidates validation and indexation. Indeed we have deployed a GIS database using Postgis[4] and a french villages layer. The validation consists then in proving candidates existence in the database. Next section also explains how GISs are used to geo-localize these validated candidates.

3.2 Geographic criterium-based information retrieval

We use information extraction to undertake queries and retrieve information from documents. For every extracted SF, an instance of the unified geographic model is created and stored in index files. This instance consists of the name of the feature, its interpretation (A_SF or R_SF with their relations) and a corresponding geometric object (representing the concerned area). During the last stage of the semantic treatment, this geometric object is computed using GIS services and our algorithm (Figure 9).

The notion of area for SF's geo-localization in index files and in queries is approached in this section. Then the spatial semantic-based information retrieval stage is explained.

3.2.1 The spatial location of geographic features

The unified geographic model can represent SFs freely from the expression mode (text, image, etc.): the common denominator is the geometric form and the geo-localization.

[4]postgis.refractions.net

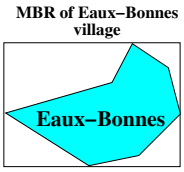

Figure 7: Eaux-Bonnes : its polygon and its MBR.

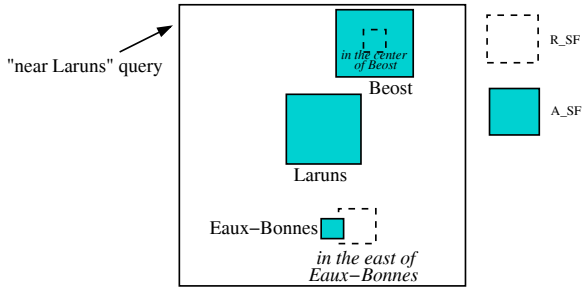

Figure 8: A query example and its matching MBRs.

Giving a representation to SFs. Every features extracted from documents are geographic data. That's why, if we want to think "area", we have to recover a geo-located representation of each SF, obviously with the help of GIS tools. If we consider the different levels of granularity and the different levels of precision, the geometrical shape corresponding to the area of a SF can change. GISs provide several geometric objects: points (a church for example) , polylines (a road), polygons, multi-polygons (a city), etc. Moreover efficients topologic functions are available in order to manage these objects.

We have developed a prototype to carry out some experiments. These experiments allow to validate our hypothesis that area-based information retrieval for corpora like those of project PIV is an efficient way of thinking. We chose to take a medium level of granularity, simplifying the complex geometrical shapes of each SF to Minimum Bounding Rectangles (MBR) (Figure 7) with a GIS topological function.

This concept of MBR has already been discussed by [19] in the Alexandria Digital Library. Moreover [11, 20] show that a MBR can be a quite good approximation of the objects' geometry. We can see an example of a query and some MBRs representing pyrenean villages in Figure 8.

Next section details the MBRs' algorithm construction. These MBRs result from the preceding semantics analyzes processes.

MBRs computing recursive algorithm. It is quite easy to ask a GIS[5] to retrieve the MBR for an A_SF (a given named place for example). However, it is more complex to retrieve the MBR for a R_SF: a GIS can not directly return the MBR for a R_SF like "*à l'est des Eaux-Bonnes*" (in the east of Eaux-Bonnes village) or "*au sud de la périphérie de Pau*" (in the south of Pau's periphery).

[5]We can also use web services to geo-localize geographic features. For example Viamichelin (http://ws.viamichelin.com/).

```
ComputeMBR(GF){
 if (GF is a AGF){
  return CallGeoRefWebService(GF);
 }
 else if (GF is a RGF){
  relation <- RelationExtraction(RGF).
  subGF <- SubGFExtraction(RGF).
  return ModifMBR(relation, ComputeMBR(subGF));
 }
}
```

Figure 9: Simplified MBRs computing recursive algorithm.

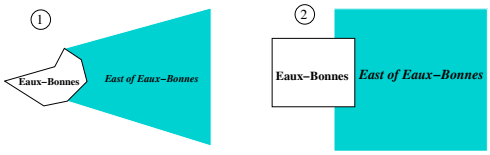

Figure 10: "east of Eaux-Bonnes village": 1/ realistic interpretation 2/ possible simplification.

To solve this problem, a recursive algorithm has been developed (Figure 9). It consists in carrying out recursively geometrical transformations on MBR. As a R_SF is always constructed using an A_SF (and one or several recursive relation(s)), our algorithm begins by retrieving the MBR of the A_SF included in the R_SF. Then, exploring recursively the relations that define the R_SF, it makes geometrical transformations on the original MBR (translations, homotheties, etc.).

Let us take an example: the MBR of the R_SF "à l'est des Eaux-Bonnes" ("in the east of Eaux-Bonnes village" Figure 10) is computed from the MBR of the A_SF "Eaux-Bonnes village", on which we carried out a translation on the x-axis. We can see on left side of the Figure 10 the "Eaux-Bonnes village" polygon and a possible interpretation of the "east of Eaux-Bonnes village". As said in the previous section, the polygon can be simplified and we obtain an including rectangle (MBR) (right side). From this rectangle representing an A_SF, we calculate an interpretation of the R_SF "East of Eaux-Bonnes village" by moving it and by extending it.

It is important to notice that for best results, this reflexion and this implementation should be applied to more complex areas like the ones given by a GIS: lines, polygons, etc.

3.2.2 *The spatial semantic-based IR process*

The search technic's principle is based on a spatial mapping between the query's SFs and the documents' SFs. This mapping is done using the MBR created dynamically for the query and stored in index files for the corpus.

Query. A query is analyzed exactly like corpus documents: the same IE data processing sequence is performed to extract every SF. In a last processing step, each extracted SF is geolocated and a MBR is attached to each one.

Index files. Each index file contains the extracted SFs with their paragraph identifiers (for text-document), their original file identifiers and their corresponding MBRs. Finally, we are able to compute the relevance of documents by simply traversing these index files.

For example, the extracted feature "east of Eaux-Bonnes village" is indexed as follows:

```
<geographic_feature id="4" id_paragraph="2">
 <r_sf>
  <label>east of Eaux-Bonnes</label>
  <relation>
   <orientation><type>east</type></orientation>
  </relation>
  <a_sf>
   <label>Eaux-Bonnes</label>
    <type>village</type>
  </a_sf>
 </r_sf>
 <presentation>
  <mbr>
   <x_min>360689.2</x_min><y_min>1752718.6</y_min>
   <x_max>389050.6</x_max><y_max>1789151.3</y_max>
  </mbr>
 </presentation>
</geographic_feature>
```

A MBR representation is added (line 13-16) to the XML-tree at the end of the semantic processing. Moreover, we use an XML database to store the index files[6].

3.2.3 *PIV information retrieval*

At this step, we do not use GIS anymore to calculate the degree of pertinence of documents but an algorithm developed in XQuery[7], which simply checks the presence of intersection between documents' MBRs and query 's MBRs.

Information retrieval's results are displayed in a Google-like presentation, that is to say a list of relevant parts of documents. Note that to test the unified spatial model with another expression mode, some pyrenean lithographs have been manually marked.

With this process, we are not only going to find the fragments of documents containing words "near Laruns" like any classical full-text research tool would have found; but we are also going to find fragments containing words "Les Eaux-Bonnes", "Beos", etc. (see Figure 8 for other examples of matching features) which are villages close to Laruns and fragments containing "à l'est des Eaux-Bonnes" ("in the east of Eaux-Bonnes village"), "au sud de Pau" ("in the south of Pau") which are R_SF also close to Laruns. The link between these features and "near Laruns" exists thanks to the semantic IE stage and this appropriate information retrieval process.

4. CONCLUSION

We focus our works on restricted corpora such as local cultural and heritage documents collections. This specific context makes possible to implement more sensitive scans that take into account the document contents.

Our contribution is complementary to traditional access methods used in digital libraries. Indeed non geographical information management is deported on existing LMS deployed by the local digital library. Our system extends this LMS for the non geographical information part of the query.

We aim at considering in a more accurately way the geographic semantics in such collections of documents and in users' queries. PIV prototype implements and combines original geographic semantics IE and IR approaches. Its experimentation with heterogeneous (texts and images) documents collections validate our unified model and shows that this approach increases geographical query results relevance.

[6]The database is called eXist: http://exist.sourceforge.net
[7]http://www.w3.org/TR/xquery/

The first experiment detailed in this paper validates our assumptions concerning the geo-localization of indexed SFs and their use during the spatial-based IR process for a regional media library corpora. We now plan to integrate more GIS tools in our system during the spatial relation management and the MBRs construction stage.

Moreover GISs tools could be used to determine other geometric objects in order to take into account various granularities. So a SF will be able to have several representations and will be defined by various geometric objects.

5. ACKNOWLEDGEMENTS

Our project is led in partnership with the community of agglomeration of Pau and its media library (called MIDR). We want to thank them for their help and their support.

6. ADDITIONAL AUTHORS

Christian Sallaberry LIUPPA - Université de Pau et des Pays de l'Adour 64013 Pau Université Cedex.
Email: `christian.sallaberry@univ-pau.fr`

7. REFERENCES

[1] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 1983.

[2] T. Baccino and J. Pynte. Spatial coding and discourse models during text reading. *Language and Cognitive Processes*, 9:143–155, 1994.

[3] J. Bertin. *La sémiologie graphique*. Gautiers Villars, 1973.

[4] C. Bessière, J. Euzenat, R. Jeansoulin, G. Ligozat, and S. Schwer. Raisonnement spatial et temporel. *Actes 6e journées nationales du PRC-GDR intelligence artificielle*, 1997.

[5] F. Bilhaut. The linguastream platform, 2003.

[6] A. Borillo. *L'espace et son expression en français*. L'essentiel. Ophrys, 1998.

[7] J. Casenave, C. Marquesuzaà, P. Dagorret, and M. Gaio. La revitalisation numérique du patrimoine littéraire territorialisé. In *Colloque international EBSI-ENSSIB, Montréal, Canada, Octobre 2004*. 2004.

[8] M. Chambreuil, editor. *Sémantiques*. Hermès, Paris, 1998. ISBN 2-86601-721-8.

[9] T. Charnois, Y. Mathet, P. Enjalbert, and F. Bilhaut. Geographic reference analysis for geographic document querying. Workshop of the NAACL-HLT Conference, Association for Computational Linguistic, 2003.

[10] M. Charolles. L'encadrement du discours : Univers, champs, domaines et espaces. Cahier de Recherche Linguistique 6, Université de Nancy2, 1997.

[11] E. Clementini, J. Sharma, and M. Egenhofer. Modeling topological spatial relations: Strategies for query processing. Computers and Graphics 18 (6): 815-822, 1994.

[12] A. G. Cohn. Qualitative spatial representation and reasoning techniques. In *KI '97: Proceedings of the 21st Annual German Conference on Artificial Intelligence*, pages 1–30, London, UK, 1997. Springer-Verlag.

[13] A. G. Cohn and S. M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1-2):1–29, 2001.

[14] M. A. Covington. Gulp 2.0: An extension of prolog for unification-based grammar, 1989.

[15] M. J. Egenhofer. Toward the semantic geospatial web. In *GIS '02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pages 1–4. ACM Press, 2002.

[16] M. E. Eliseo Clementini, Jayant Sharma. Modeling topological spatial relations: Strategies for query processing. *Computers and Graphics*, 1994.

[17] P. Enjalbert and M. Gaio. Traitements sémantiques pour linformation géographique, textes et cartes. *Geomatique*, 2005. to be published.

[18] P. Etcheverry, C. Marquesuzaà, and J. Lesbegueries. Revitalisation de documents territorialisés : Principes, outils et premiers résultats. Workshop Met-SI INFORSID, 2005.

[19] L. L. Hill. Core elements of digital gazetteers: Placenames, categories, and footprints. In *ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, pages 280–290. Springer-Verlag, 2000.

[20] Larson and Frontiera. Ranking and representation for geographic information retrieval. Workshop on Geographic Information Retrieval - SIGIR, 2004.

[21] P. Loustau. Traitements sémantiques de documents dans leur composante spatiale. Master's thesis, Université de Pau et des Pays de l'Adour FRANCE, 2005.

[22] N. Malandain, M. Gaio, and J. Madelaine. Improving retrieval effectiveness by automatically creating some multiscaled links between text and pictures. In *Proceedings of SPIE, Document Recognition and Retrieval VIII*, volume 4307, pages 89–99, San Jose, Californie, USA, 24-25 Janvier 2001.

[23] C. Marquesuzaà, P. Etcheverry, and J. Lesbegueries. Lecture notes in computer science. volume 3799, chapter Exploiting Geospatial Markers to Explore and Resocialize Localized Documents, pages 153–165. Springer-Verlag GmbH, November 2005.

[24] D. W. Oard and G. Marchionini. A conceptual framework for text filtering process. Technical Report CS-TR-3643, 1996.

[25] M. Torres. Semantics definition to represent spatial data. International Workshop - Semantic Processing of Spatial Data - Geopro, 2002.

[26] C. Vandeloise. *L'espace en français*. Travaux Linguistiques. Seuil, 1986.

[27] A. Widlocher and F. Bilhaut. La plate-forme linguastream : un outil d'exploration linguistique sur corpus. In *Actes de la 12e Conférence Traitement Automatique du Langage Naturel*, 2005.

[28] A. Widlöcher, E. Faurot, and F. Bilhaut. Multimodal indexation of contrastive structures in geographical documents. In *Actes RIAO 2004, Avignon*, pages p. 555–570, 2004.

[29] M. Worring, A. D. Bagdanov, J. van Gemert, J.-M. Geusebroek, H. Minh, G. Schreiber, C. Snoek, J. Vendrig, J. Wielemaker, and A. W. M. Smeulders. Interactive indexing and retrieval of multimedia content. In *SOFSEM*, pages 135–148, 2002.