Accessing Heritage Documents according to Space Criteria within Digital Libraries

Christophe Marquesuzaà, Patrick Etcheverry, Christian Sallaberry, Mustapha Baziz

Accessing Heritage Documents according to Space Criteria within Digital Libraries

Christophe Marquesuzaà[1], Patrick Etcheverry[1], Christian Sallaberry[2], Mustapha Baziz[2]
[1] LIUPPA
IUT de Bayonne
Château-Neuf, Place Paul Bert
64100 Bayonne - FRANCE
Patrick.Etcheverry@iutbayonne.univ-pau.fr
Christophe.Marquesuzaa@iutbayonne.univ-pau.fr
[2] LIUPPA
Université de Pau et des Pays de l'Adour
Avenue de l'Université, B.P. 1155
64013 Pau Cedex - FRANCE
Christian.Sallaberry@univ-pau.fr
Mustapha.Baziz@univ-pau.fr

Abstract: *Local cultural heritage document collections are characterized by contents strongly attached to a territory and its associated land history. Our contribution aims at enhancing such a content retrieval process efficiently each time a query includes geographic criteria. We propose a unified model for a formal representation of geographic information. This geographic model allows space features to be described independently of their representation mode (text, graphics) in the documents. We have developed a prototype implementing geographic Information Extraction (IE) and geographic Information Retrieval (IR) processes. We process geographic IE with semantic techniques combined to classic IE approaches. Then, we implement geographic IR with intersections researching algorithms: these algorithms search for all geocoded entities in the documents collections indexes which intersect any entity in the user's query. This paper focuses on IR and Visualization proposals relying on the geospatial characteristics of cultural heritage corpora.*

Keywords: Geographic Model, Geographic Information Retrieval and Visualization, Non-Structured Documents, Digital Libraries, Cultural Heritage

1. Introduction

Smart space information retrieval and visualization is the main goal of the work presented in this paper. Generally, space information is either supported by RDBMS (Relational Data Base Management Systems) and GIS (Geographic Information Systems) 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. Although GISs contain high-level space operators that are uncommon in conventional DBMSs, they are not sufficient for queries in which the semantics of the search criteria concerns space relations [Clementini & al., 1994]. The results are also unsatisfying if we consider EDMSs that usually implement statistical approaches to answer such queries.

The purpose of the Virtual Itineraries in the Pyrenees (PIV – Pyrénées Itinéraires Virtuels) project consists in managing a repository of electronic versions of books, newspapers, postcards, lithographs of the XIXth and XXth century. Information is mainly textual and presents many territorial aspects of the Pyrenees (a mountain range in the south west of France) [Casenave & al., 2004]. This corpus is still relatively unknown. It is accessible only in regional museums and library archives. This is why the local media library supporting this project aims at the diffusion of these resource collections: their added-value remains centred on local cultural heritage and, therefore, geographic characteristics. To complete statistical and full-text analysis approaches, we propose a more accurate semantic approach to analyze and interpret geographic information contained in such a corpus (or in a query) [Marquesuzaà & al., 2005], [Etcheverry & al., 2005], [Sallaberry & al., 2006].

Geographically related queries form nearly one fifth of all queries submitted to the Excite search engine, the terms occurring most frequently being place names [Sanderson and Kohler, 2004]. Our contribution focuses on digital libraries and proposes to extend the basic services of existing Library Management System with new ones dedicated to geographic information extraction and retrieval (PIV project [Lesbegueries & al., 2006]).

Our contribution aims at better integrating the space dimension in the information retrieval systems. In a more accurate way, our contribution focuses on five proposals:

- we both propose a space unified model allowing any space feature to be described as well as an indexing process adapted to space features presented within analyzed documents.
- we describe how to use this indexing process in order to select documents in a more accurate way basing on the geographical areas described in these documents.
- we specify the various criteria that can be considered in order to visualize the resulting documents of a space query in a more relevant way.
- we concretely illustrate these proposals with a prototype implementing these main ideas.
- we detail our first evaluation results concerning both space information extraction and retrieval processes.

We present related works in the next section and an overview of the PIV system in the third section. PIV space content-based information extraction, retrieval and visualization are respectively presented in sections four, five

and six. For each chapter, we first present the related works, we then define the main principles and associated concepts in order to detail our proposal with both a conceptual and a technical aspect.
Section 7 discusses the PIV project evaluation first results and is followed by a section that is dedicated to conclusion and future work.

2. Space Information Management within Heterogeneous Documents Collections

Information Extraction (IE) generally organizes indexes for a better support Information Retrieval (IR). IE and IR used together have the potential to create powerful new tools in information processing [Gaizauskas, 2002] [Gaizauskas and Wilks, 1998]. This section describes the IE, IR and Information Visualization (IV) approaches which are combined for specific geographic requirements [Sallaberry & al., 2006].

2.1 Information Extraction

IE activity aims at building a structured information repository from an unstructured information source [Gaizauskas, 2002]. In a collection of documents, the result of IE constitutes what is called an index. It is generally made up of a list of terms linked to each document [Tebri, 2004]. These terms have to describe the contents of the documents as precisely as possible. Automatic IE processes extract either the whole information in a document or, specific parts of it. For example, in the first case, text processes generally use statistical approaches (each term of a document is processed) [Zipf, 1949] [Tebri, 2004] to associate a weight to each term while, in the second one, they use predefined rules in order to find specific information [Gaizauskas, 2002].

2.2 Information Retrieval

Information Retrieval (IR) deals with models, techniques, procedures to extract information that has already been processed, organized and stored (databases, files, XML files, etc.). [Baeza-Yates and Ribeiro-Neto, 1999] explain that satisfying the user information requirements is not trivial: "*The user first specifies a user need which is then parsed and transformed. Then, query operations might be applied before the actual query, which provides a system representation for the user need, is generated. The query is then processed to obtain the retrieved documents. Fast query processing is made possible by the index structure previously built.*". [Baeza-Yates and Ribeiro-Neto, 1999] and [Torres and Parkes, 2000] stress the importance of the query validation and/or reformulation stages to improve the interpretation of the user need.

2.3 Information Visualization

The retrieval of search engine results is a recurring problem because it is important for the user to make use of and to efficiently visualize the retrieved information [Bonnel and Moreau, 2005]. We want to highlight that the restitution step is always coupled with a calculation and a representation step of the information relevance (according to relevance calculation methods, restored information can differ for the same need expressed). Here is a summary of the different retrieval modes that presents the four most used approaches:

- *List representation*: A list presents a collection of elements organized according to a relevance mode defined by the list manager. It is a representation method used by most search engines like Google[1] or Grokker[2]. [Hascoët and Beaudoin-Lafon, 2001] present a taxonomy of the information visualization strategies relying on techniques adapted to large size lists (*e.g.* the perspective wall technique of Mackinlay [Macinlay & al., 1991]). The relevance of an element is represented by its position in the list. Lists allow to present indefinite result numbers and also provide a simple access mode to the items. However, they do not offer a synthetic overview of the results and the browsing is often limited to the first results [Leuski and Allan, 2000] [Bonnel and Moreau, 2005].
- *Topic representation*: A topic gathers the elements around same concept. The Ujiko[3] or Kartoo[4] search engines use this type of representation. [Kules and Schneiderman, 2004] showed that, in comparison with a non-hierarchical classification, such a classification could give better performances in terms of user satisfaction. This representation offers a synthetic overview but choosing the topics to be represented is subjective and difficult to computerize. The relevance is not explicitly represented because it depends on the matching degree between the presented topics and the user requirements.
- *Graph or Tree representation*: It is close to topic classification but it integrates the semantics carried by the edges connecting the vertices that can represent topics or documents. The Kartoo search engine or topic maps [Cossanel & al., 2002] with tools such as TM4J[5] use this form of representation. The semantic links between concepts allow guided navigation between topics and documents. However, there are two limits to the number of concepts that can be represented: the size of the display for results presentation and the risk of cognitive overload for the user. As for topic representation, the relevance is not explicitly represented, it depends on the matching between the user requirements and the suggested thematic browsing links.
- *Cartographic representation*: The principle consists of representing a space (area, country, building, room, device layout) and associating the data elements related to various points highlighted on this space. An example of geographical application is SPIRIT - Spatially-Aware Information Retrieval one the Internet[6] - which proposes a search engine whose results are Web pages geo-localized on a map. With this approach, the user

[1] www.google.com
[2] www.grokker.com
[3] www.ujiko.com
[4] www.kartoo.com
[5] http://tm4j.org/
[6] www.geo-spirit.org

reads the map looking for space reference marks; the relevance is thus spatially represented by the proximity of results with a space reference point of interest for the user. This form of representation integrates the space dimension but raises problems when the results are geo-localized on very close places, or on the same place.

- *2D-3D representation*: A particular case of visualization modes deals with element representation in two or three dimensions. All the previous representation modes can be modeled either in 2D or 3D. 3D representations are frequent when the amount of documents is significant [Jacquemin & al., 2005]. The choice of a 3D representation is justified by the need to increase the visualization space [Bonnel & al., 2005a] [Bonnel & al., 2005b]. However, the third dimension entails a more complex navigation.

Geographic contents management within heterogeneous documents collections is the main purpose of the paper. We now present the linguistic and GIS approaches to space information management.

2.4 Space Information

- *Linguists' works*: They explain our specific manner of representing space information in written language. According to [Borillo, 1998], we can link a place to a category and associate it with a natural or artificial boundary. We consider three main categories: plots of land, expanses of water, dwelling places. Referring to such places involves several elements. In written language, one might define space information by referring to a better known position. We thus understand sentence 1 perfectly while sentence 2 sounds unusual to us:
  - sentence 1: "*the car is near the house*"
  - sentence 2: "*the house is near the car*"

  [Vandeloise, 1986] studied this assumption for textual documents and explained the concept of the target/site couple. Our objective is to extend this hypothesis to any other expression modes.
- *GIS works*: They present a Space Feature (SF) as a user-defined geographic phenomenon that can be modeled or represented using geographic data sets. Examples of geographic features include streets, sewer lines, manhole covers, accidents, lot lines, parcels[7]. Important related works address models of space relations [Clementini & al., 1994], qualitative space representation and reasoning [Cohn, 1997] [Cohn and Hazarika, 2001] [Muller, 2002] and space queries processing [Clementini & al., 1994]. Other interesting works [Hill, 1999] [Hill, 2000] concern digital gazetteers (Alexandria Digital Library[8]) which can be defined as geo-spatial [Zahn, 2001] and support important related dictionaries of geographic names. GIS literature mainly represents a geographic feature by its name and location. Location covers a large amount of different concepts:
  - topographical coordinates, with geometric possibilities (the point or the polygon coordinates to locate the building on a map);
  - topological, direction or metric relations with other SFs (a direction relation to detail the position of the building within a village) [Bonnel and Moreau, 2005] [Egenhofer, 2002] [Gotts and Goodday, 1996];
  - conceptual links with topics of space theory within a specific ontology [Cohn and Azarika, 2001] [Hernandez, 2005].

As IE, IR and IV approaches are rather generic, space information accurate management still represents a great challenge. Moreover, Library Management Systems can neither take into account the geographic semantics of heterogeneous documents nor the users' specific geographic requirements. Semantic processing seems to be an interesting way of space information management within IE, IR and IV.

2.5 Semantic Processing

It allows specific information extraction; *i.e.*, exploiting the localized property of a corpus in order to focus on the space information. In textual expression mode, the data processing sequence used for highlighting space markers is composed of four main steps [Abolhassini & al., 2003]:

- lemmatization carries out a segmentation of the words;
- lexical and morphological analysis proceeds to a word recognition;
- syntactic analysis, based on grammars, allows to find the links between words;
- "semantic" analysis carries out a more specific analysis allowing the extracted syntagms to be interpreted.

Some systems like Brill[9], Cordial[10] or Tree-Tagger[11] morphosyntactical analysers are dedicated to a specific part of such sub-processes. Other systems like Linguastream[12] [Widlocher and Bilhaut, 2005], SPIRIT[13] or GATE[14] [Gaizauskas & al., 1995] [Gaizauskas, 2002] support the whole process.

In the graphic expression mode, semantic processes consider that an image is not represented in single pixels but in meaningful image segments and their mutual relations. [Torres, 2002] and [Enjalbert and Gaio, 2006] propose a semantic definition to represent space data. [Benz & al., 2002] present fuzzy methods implementing expert space knowledge and describe a workflow from remote sensing imagery to GIS. The eCognition system provides a new technology for image analysis[15].

---

[7] www.webgis.net/cms.php/glossary.htm
[8] www.alexandria.ucsb.edu; www.alexandria.ucsb.edu/gazetteer/
[9] www.cs.jhu.edu/brill/
[10] www.synapse.com
[11] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[12] www.linguastream.org
[13] www.spiritengine.com
[14] http://gate.ac.uk/
[15] www.pcigeomatics.com/products/definiens.html

3. Overview of the PIV System

3.1 Information Visualization Overview

The PIV prototype is a system focused on smart space information retrieval and visualization. With this system the user can seek documents referring to a particular place or geographical area.
The user query is expressed in natural language with geographical criteria which can be more or less precise: it can deal with seeking all documents describing a particular place (the city of Pau for example) or a more or less defined geographical area, for example (Fig. 1-top), seeking all the documents referring to the area between Pau and Lourdes cities.
Since information retrieval is operated on space criteria, the PIV system restores its results using the same approach: a map of the area evoked in the user query is displayed and the resulting documents are presented as icons close to the places they evoke (Fig. 1).

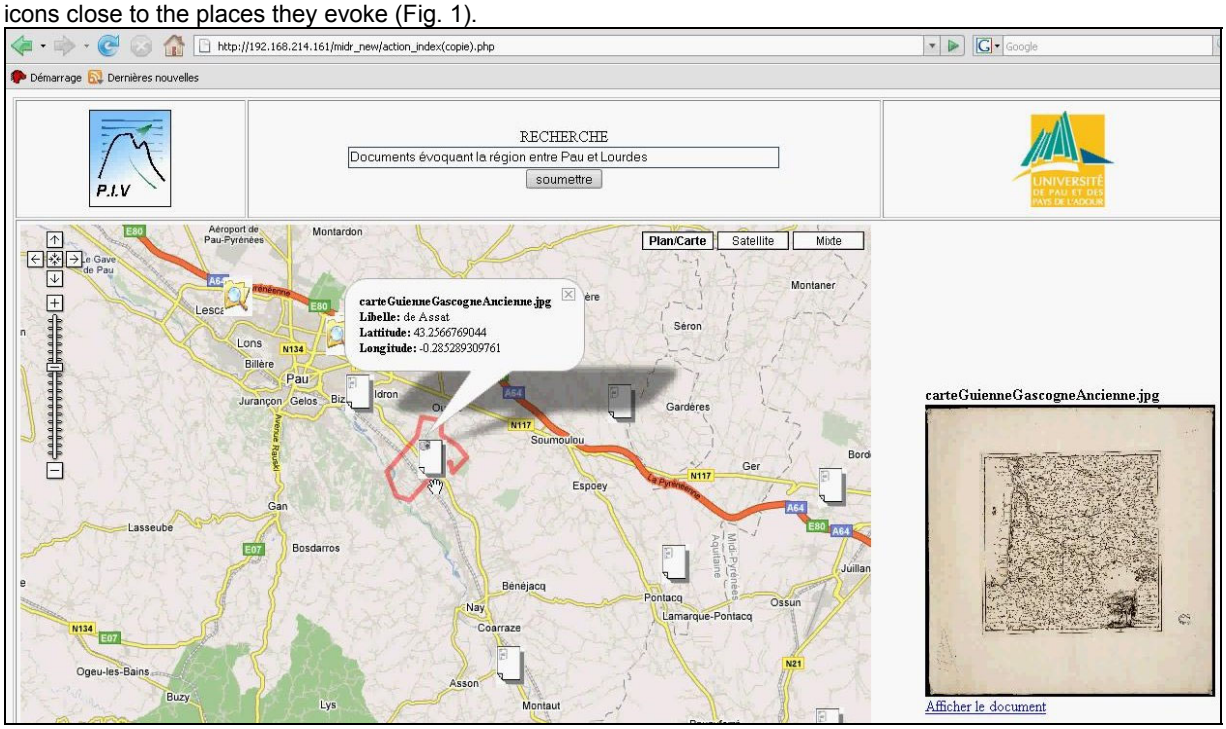

Figure 1: The PIV system – Exploring the results

At this stage, the user has an overview of the results according to a visualization mode which takes into account its geographical interests. The user can evaluate the number of resulting documents, the kind of documents (texts, iconographic documents, etc. are presented by different icons) and especially the areas described by these documents. When the user points at a document, the system presents the related geographical information in an information-bubble, as well as a thumbnail of the document (right side of the screen). If the user considers the presented information as relevant, he/she can access and consult the original document. On the contrary, he/she can explore the other documents results.
To assist the user in exploring the set of results, the system can display additional information for each document (Fig. 1). This information is mainly of two kinds: geographical and documentary. Thus, on a document mouse over, the system provides additional information concerning the area evoked in this document (area displayed on the map, geographical co-ordinates etc.). With a mouse click on the document, the system provides an overview of the document itself. This overview takes the shape:

- of a thumbnail if it is about a graphic document.
- of a paragraph if it is about a text document. The selected paragraph always evokes the space features which justify why this document was considered to be relevant compared to the initial query.

3.2 Information Extraction and Information Retrieval Overview

Fig. 2 presents the PIV IE and IR processes. The IE process is composed of four main stages:

- First of all (stage (1)), each document is OCRized and classified according to its nature (text, image, etc.).
- In stage (2), the linguistic and semantic analysis is carried out in order to extract potential space features (SFs). Each potential SF is formally described according to a unified space model described in section 4.3.1.
- The third stage (3) parses geographic gazetteers (districts, named-places, roads, cliffs, valleys, …) in order to validate each SF detected in the previous stage.
- In stage (4), the IE process computes the space representation and the georeferences of each validated SF. This information (instance of the space model and georeferences) is used to index each document.

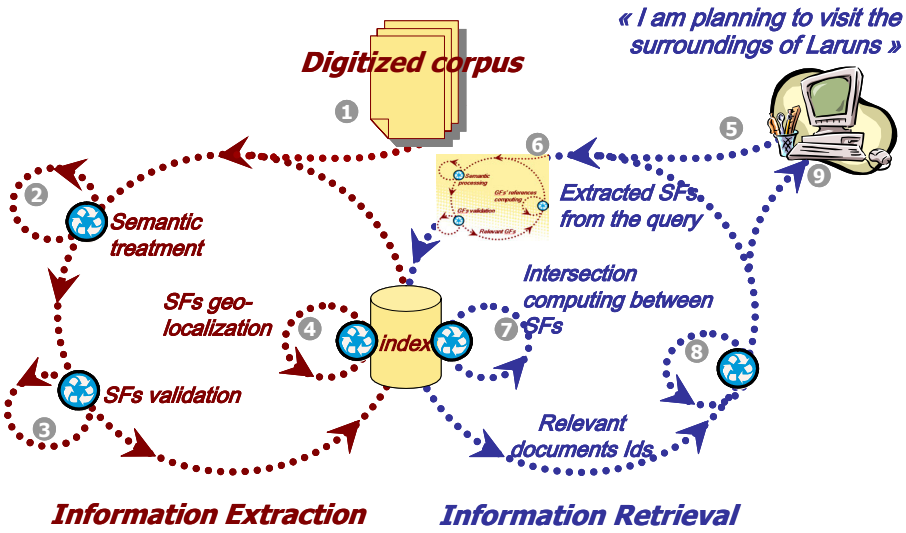

Figure 2: The PIV system – Information Extraction and Retrieval processes

As the user query (stage (5)) is expressed in text form, this query may be considered as a textual document. Hence, the query may be processed (stage 6) with the same previous IE process. After this stage, each SF evoked in the user query is also spatially described as previously and in accordance with our space model. In stage 7, this same representation therefore allows to compute intersection surfaces between the space representations evoked in the query and in the indexes. Stage 8 selects the corresponding relevant documents which are presented to the user in stage (9). This last stage corresponds to the Information Visualization (IV) part of the PIV project.

We will now detail in a more accurate way how the PIV system implements IE, IR and IV approaches to better manage space information in a cultural heritage corpus.

4. Information Extraction

4.1 Related Works

Most Information Extraction (IE) systems use statistical approaches. They are dedicated to large and repeatedly revised corpora (*i.e.* web). Slight statistic processing chains operate re-indexing processes. [Chen & al., 2006] give an overview of indexing research works related to thematic and/or geographic information within web corpora. In the case of pure thematic information extraction, the most used methods are statistics-based and weigh the extracted terms according to both local and global term frequencies [Salton and McGill, 1984] [Robertson, 1977]. The idea behind the document term local frequency (TF) is to estimate the importance of a term in the document; whereas the document term global frequency (also known as inverse document frequency or IDF) is used to estimate the discrimination power of the term according to the whole collection. Let us note that in most statistical IE approaches, the extracted terms are first stemmed according to [Porter, 1980] in order to reduce the length of the index and then stop-words are removed according to a standard list [Salton and McGill, 1984].

Other IE methods aim at Named Entity Recognition (NER). They consist in assigning a word or a word group to a set of standard entity types to extract: organization, person, location, date, time, money and percent entity types. The most straightforward means of identifying named entities is through matching document contents to previously generated lists of names. Another approach aims to incorporate context characteristics thanks to NE grammars in order to better capture external evidence. Rules for NE recognition can be generated entirely by hand (knowledge-based) or automatically using machine learning (or statistical) techniques. More sophisticated methods of NER make use of both lists and NE grammars. For instance, LaSIE[16] is an integrated information extraction system designed for research and benchmarking purposes. It supports the TRESTLE system [Gaizauskas, 2002].

Some IE systems complete NER approaches with an additional semantic processing stage. They are dedicated to domain specific, stable and quite small corpora. As there is no or very little revision, a hard back-office semantic process seems to be suitable. [Lesbegueries & al., 2006] describe a geographic information markup process dedicated to local cultural heritage collections of documents. [Wildöcher & al., 2004] present multimodal indexing in geographical documents.

The last step concerns indexing. For text information, [Baeza-Yates and Ribeiro-Neto, 1999] promote two main indexing techniques: inverted files and suffix array. They emphasize inverted files which are currently the best choice for most keyword-based search applications. Suffix trees and arrays are faster for phrase searches and other less common queries, but are more difficult to implement and to maintain.

[16] www.dcs.shef.ac.uk/research/groups/nlp/funded/lasie.html, www.destarter.com/lassie/lasie.html

4.2 Space Information Extraction Related Works

Most NER systems dedicated to space information management consist of at least three basic components: (1) a tokenizer, (2) gazetteer lists, and (3) a NE grammar. The tokenizer segments text into tokens, *e.g.* words, numbers, and punctuation. The gazetteer contains lists of named entities, *e.g.* towns, countries, cities, and lists of keywords such as titles and company designators (*e.g.* Plc and Ltd). The NE grammar consists of rules for NE recognition, which take the context into account.

The SPIRIT project [Jones & al., 2004] proposes a method of NER using gazetteer lists in combination with grammars. The SPIRIT engine parses semi-structured web page texts looking for occurrences of text addresses and zip codes, place names and telephone numbers… So, a web page is annotated with a geographic footprint and a full-text index [Vaid & al., 2005]. In the same way, [Sagara and Kitsuregawa, 2004] use Yellow Pages to generate key words to find documents on the web related to listed businesses. The SPIRIT project and other works like the Geosearch system, the GEO-IR system, etc. are related to space information management and are presented in [Chen & al., 2006]. The IE process last step concerns indexing; [Baeza-Yates and Ribeiro-Neto, 1999] propose the R-tree indexing method in order to represent a space object by its Minimum Bounding Rectangle (MBR). Data rectangles are grouped to form parent nodes, which are recursively grouped to form grandparent nodes and, eventually, a tree hierarchy. Another way of indexing text and/or multimedia documents is the ontological approach [Ihadjadene & al., 2004a] [Ihadjadene & al., 2004b]. With this approach, a document or a part of document can be marked up referring to a specific topic (space for footprints computation, architectural for historical period or artistic class deduction, etc.). A drawback of such metadata is that it is usually built manually, sometimes semi-automatically. Most of the time, ontologies are used to extend a query criteria rather than document indexing.

4.3 Space Information Extraction within the PIV Prototype

We use a tokenizer, a NER method with gazetteers and a Definite Clause Grammar (DCG) but we extend this analysis process with an additional semantic processing stage supported by a space core model aiming at Space Features (SF) interpretation [Lesbegueries & al., 2006]. An SF might be a simple (absolute) named entity or a more complex (relative) one (*cf.* section 4.3.1).

A difference of our approach with others like SPIRIT [Jones & al., 2004] and GIPSY [Woodruff & al., 1994] relies on the back-office space reasoning mechanims used for both absolute SFs (ASFs) and relative SFs (RSFs) interpretation and indexing. For instance, the SPIRIT system mainly tags ASFs.

Next, indexing principles are only applied on textual resources. The indexing of non textual resources (lithographies, postcards, etc.) is carried out on textual metadata provided by the MIDR library management system. This metadata corresponds to the usual index cards manually produced by the MIDR archivists.

Other specificity concerns the granularity level of the managed information units: text paragraphs of our domain specific corpora (the cultural heritage of the Pyrenees) and web pages in the case of the SPIRIT system.

Moreover, in our proposal, the same refined space information markup and interpretation process is applied both in the information units indexing stage and in the users' query interpretation.

4.3.1 Conceptual Proposal

The PIV project information extraction (IE) process aims at locating, marking and describing (within an index) the SFs mentioned in each document of the corpus. To complete this work, we proposed [Lesbegueries & al., 2006] a space model (Fig. 3 top right corner) describing any SF which can be either:

- an Absolute Space Feature (ASF) if it only consists in a named entity allowing geo-localization, or
- a Relative Space Feature (RSF) if it is defined using a topological relation with at least one SF.

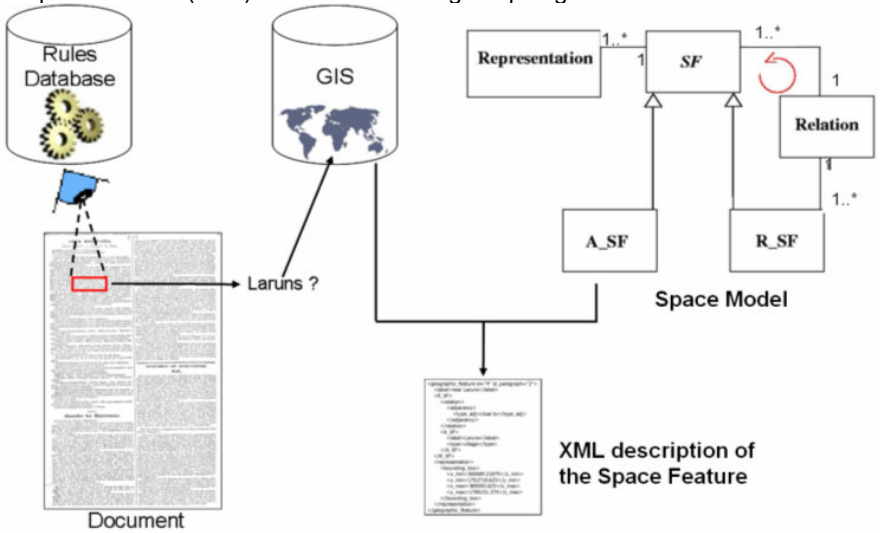

Figure 3: Space indexing of the corpus

Space features such as "*Biarritz city*" are well-known named places and are considered as ASF. Complex Space Features like "*Biarritz city*" or "*South of Biarritz*" have to be interpreted and, therefore, require some space reasoning processes [Cohn and Hazarika, 2001]). Such features are considered as RSF. We associate each RSF to one or more topological relationships for a recursive definition. Topological relations can be adjacency, inclusion, distance, geometric form and orientation [Cohn and Hazarika, 2001], [Covington, 1994].
This allows to describe ASFs ("*Laruns village*") as well as RSFs ("*close to Laruns*", "*in the south of Laruns*"). The space IE process is then carried out automatically by a set of grammar rules [Lowgren and Stolterman, 2004] [Lesbegueries & al., 2006] which seek text patterns that may refer to a SF. Any potential SF is then compared with geographical resources and any validated SF is described within indexes in accordance with the space model (Fig. 3).
Among descriptive information, we notably keep the wording of every SF (*i.e.* "*close to Laruns*"), the kind of relationship (*i.e.* adjacency) and its space representations computed with a GIS according to the topological relationships interpretation algorithms [Lesbegueries & al., 2006]. SFs extracted from various expression modes are formally represented as instances of the unified model. The resulting index files are an XML and a GIS files as shown in Fig. 4-a and 4-b.

```
<geographic_feature id="4" id_paragraph="2">
  <label>near Laruns</label>
  <R_SF>
    <relation>
      <adjacency>
        <type_adj>close to</type_adj>
      </adjacency>
    </relation>
    <A_SF>
      <label>Laruns</label>
      <type>village</type>
    </A_SF>
  </R_SF>
  <representation>
    <bounding_box>
      <x_min>360689.21875</x_min>
      <y_min>1752718.625</y_min>
      <x_max>389050.625</x_max>
      <y_max>1789151.375</y_max>
    </bounding_box>
  </representation>
</geographic_feature>
```

doc	SF_id	Paragr_id	...	SF_type	SF_geom	...
12	4	2	...	D	360689...	...

Figure 4: (a) a slight XML description of a SF, (b) the GIS description

4.3.2 Technical Proposal

The PIV prototype is built on a flexible architecture based on web services in order to facilitate their combination or the integration of new ones and therefor, the evolution of the application. The IE web services are combined into three main modules: a linguistic/semantic processing module, a representation computing module, and an indexing module. Any of these web services may be invoked by IE applications.
Therefore, the PIV back-office process annotates every document progressively: web services are invoked successively to markup the document with specific space information quoted by XML tags (Fig. 4-a). The document analysis, aiming at identifying the candidate SFs, is carried out with the Linguastream[17] tool. Some Linguastream analysis modules have been transformed into web services [Marquesuzaà & al., 2005] in order to be more easily integrated to our system. The validation and representation of the SFs identified by web services are carried out using space resources, interpretation algorithms and the PostGIS[18] geographical information system. Resources contain a geographical description (Lambert II format) of areas and places of the French Pyrenean territory. The description of the SFs is performed in an XML format and is stored both in an eXist[19] index database and a PostGIS index database (Fig. 4-a and 4-b).

5. Information Retrieval

5.1 Related Works

In a classic IR approach, a weight is assigned to each term in a document. [Robertson & al., 1998] introduced an enhanced TF.IDF formula to attenuate the negative impact of large documents in the searching stage. The same indexing process is applied to queries. A vector-based model is then used to retrieve documents: for a given query, the similarity between the vector of the query and the vectors of each document in the collection is carried out by computing the inner product between the vectors [Boughanem & al., 2001]. This relevance score is used to determine the ranking of the document in the final list of retrieved documents in response to the query.

5.2 Space Information Retrieval Related Works

A space IR involves indexes that represent every item of space information with one or more space geometries [Bressan & al., 2000] or hierarchies of quadrilateral cells defined by latitude and longitude [McCurley, 2001]. Queries ask for documents which match some keywords or contain addresses within a given radius of a specified target place.

[17] www.linguastream.org
[18] postgis.org
[19] exist.sourceforge.net

The SPIRIT system is a research product combining both thematic and space IR [Vaid & al., 2005]. The SPIRIT engine uses two kinds of ontologies: one dedicated to locations and basic geographic relations (north, south, etc.) and another one dedicated to business types within a location (*e.g.* hotels, train stations, touristic sites, etc). For instance, the SPIRIT system proposes an approach integrating space indexing with textual indexing by means of spatio-textual keys [Jones & al., 2004]: each document is associated with a geometric footprint and a set of significant textual terms. Each footprint is associated with space cells with which it intersects (a space-directed indexing method using a regular grid) and each term is associated with a list of the documents (inverted document list) in which it occurs.
Therefore, an element of a SPIRIT spatio-textual index looks like:

`"spirit" – {R1(D1;D7);R2(D3;D11;D13);R3(D2);R4(D8;D9;D11)}.`

Here, "`spirit`" is the indexed term, `D = {D1,D2, …}` is a collection of documents, `R = {R1,R2, …}` is the document space divided into space cells. This means that the "`spirit`" term is employed in documents `D1` and `D7` when associated to the `R1` space area.
Such a structure is exploited to first search for a textual term and then using the associated space index of documents to filter out those meeting the space constraints [Jones & al., 2004].
Such a spatio-textual index parsing looks like pure text index (inverted files) parsing. In the initial prototype, the SPIRIT space relevance computing is based on measures of distance between the query footprint and the document footprint [Vaid & al., 2005].

5.3 Space Information Retrieval within the PIV Prototype

As the PIV context concerns local cultural heritage documents, our approach proposes to consider space information as the most stable element of PIV's geographic information; it might be the first research criteria, the temporal information could be the second one and the thematic the third one.

5.3.1 Conceptual Proposal

A free-text querying interface supports the IR stage (IR part, Fig. 2). Any query or document in the corpus is analyzed with the same process: the IE sequence is processed and the SFs of a query are extracted. Then, the validated SFs are geo-localized and a geometry is attached to each one of these SFs.
Our search technique is based on a space mapping between the query's SFs and the documents' SFs. This mapping is done thanks to the geometries created dynamically for the query and the geographic representations stored in the index files of the corpus.

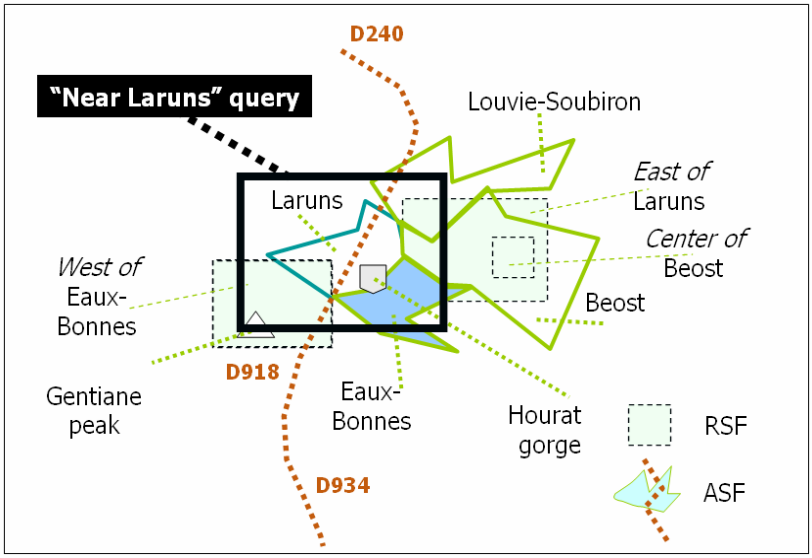

Figure 5: Example of a query and some indexed Pyrenean villages and roads

For example, Fig. 5 illustrates a query ("*I want documents dealing with places which are near Laruns.*") and its corresponding Multiple Bounding Rectangle (the biggest MBR). The other shapes represent SFs (extracted from the documents in our corpus) that may match the query. Indexed document RSFs are represented by MBRs whereas ASFs are represented by more precise geometric shapes. The relevance of a document is computed from the intersecting surface of its SF and that of the query.
We are therefore able to calculate the relevance of documents by simply crossing index files and computing SFs representations intersecting surface rates and their centroid distances [Sallaberry & al., 2007b].

5.3.2 Technical Proposal

An XML DBMS (eXist) and a GIS (PostGIS) support these searching and relevance computing operations on the corpus indexes. Fig. 6 illustrates relevance computing via functions and queries submitted to the GIS.

area(intersection(Q_geom, Df_geom))	I_surface
area(Df_geom)	Df_surface
distance(centroid(Q_geom), centroid(Df_geom))	d
distance(centroid(Q_geom), geomfromtext('corner coordinate'))	D

```
SELECT pi.gid, pi.doc_name, pi.par_id, pi.SF-name,
(tq.isurf/tq.dfsurf + tq.isurf/tq.qsurf)/(2 + tq.d/tq.D)
AS weight
FROM piv_index pi, temp_query tq
WHERE pi.gid=tq.gid ORDER BY weight DESC;
```

Figure 6: Surfaces, distances and score computing.

The query in Fig. 6 returns the relevant documents and paragraphs identifiers. Then the original texts and the SFs details may be presented in a weighed order. A set of web services implement three main modules. A first one parses the GIS indexes and selects all the relevant SFs. Then, the second module calls score computing (cf. Fig. 6) web services in order to associate a relevance score to each of the selected SFs. A third module proposes XML index parsing services in order to retrieve more details about the SFs (text, relationships, toponymic resource (city, river, mountain, etc.). This third module is dedicated to the information visualization stage described hereafter.

6. Information Visualization

6.1 Related Works

The corpus we consider is strongly localized as defined in chapter 2.3. We thus use the cartographic representation mode in our PIV prototype.

Cartography can be defined as a set of techniques aiming at the production of maps. It is used to support the presentation of information related to space in a clear and pleasant way. Its principle consists in representing a space (area, country, building, room, device layout) and associating to it information elements related to various points placed on this space. For this representation mode, the position of an element on the map as well as the distance between two elements on this map has a significant meaning for the interpretation which will be made by a user.

Some search engines use this type of representation. Kartoo is a meta-search engine which synthesizes the results of several search engines (Fig. 7).

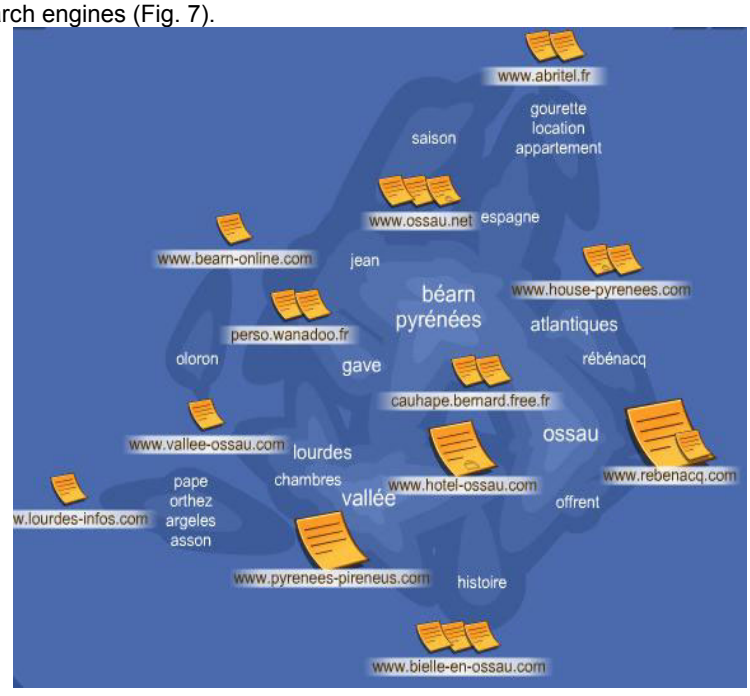

Figure 7: Kartoo metasearch engine

It presents the result of a search in the form of a map. The found results, *i.e.* web sites, are represented by more or less large pages, according to their relevance.

SPIRIT is a European project in collaboration with the French IGN[20] which aims at building a geographical search engine able to interpret the geographical elements of a query and to answer it by geo-localized documents.

[20] Institut Géographique National – National Geographic Institute – www.ign.fr

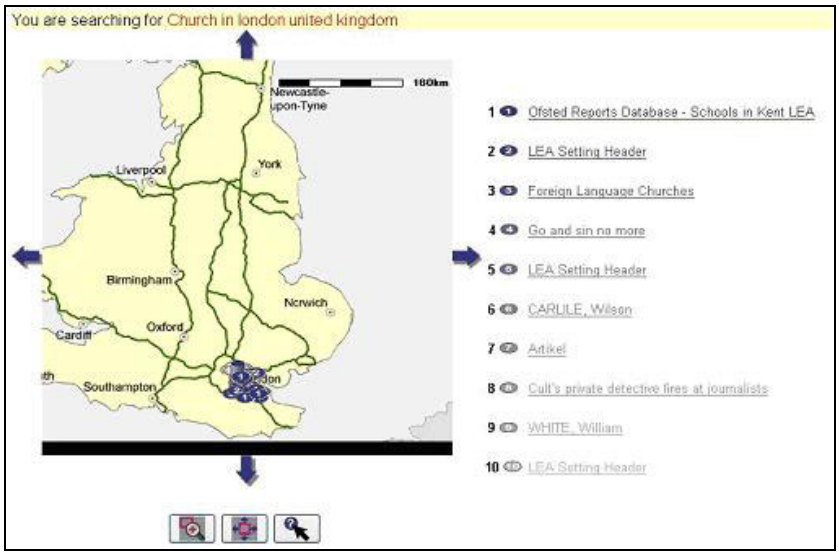

Figure 8: SPIRIT screenshot

Fig. 8 presents the interface of the SPIRIT prototype. The search results are both presented on a geographical map and on a list. On the map, different colours and sizes allow to represent the relevance of a document compared to another.

6.2 Space Information Visualization Related Works

The map representation mode is the most appropriate to design our visualization system. We have deeply studied the SPIRIT project which is the closest to our purpose. We have then defined its limits in order to propose solutions.

Although it is intuitive and attractive for a user, the presentation of documents on a map brings out various problems which must be identified and solved. Hereafter, we first present the identified problems with the cartographic representation mode.

6.2.1 Representation of documents evoking the same place (stacks)

Presenting the documents according to the places they evoke is very adequate for the user interested in a specific area or in several different places; the documents are then distributed on the map and the user can therefore start navigation by first selecting one place to explore.

The cartographic representation mode can however set a problem when the space discriminator which allows to distinguish two documents is null: all the documents evoking the same place will be represented at the same place, thus creating a "stack" of documents geo-localized on a single place (Fig. 9a).

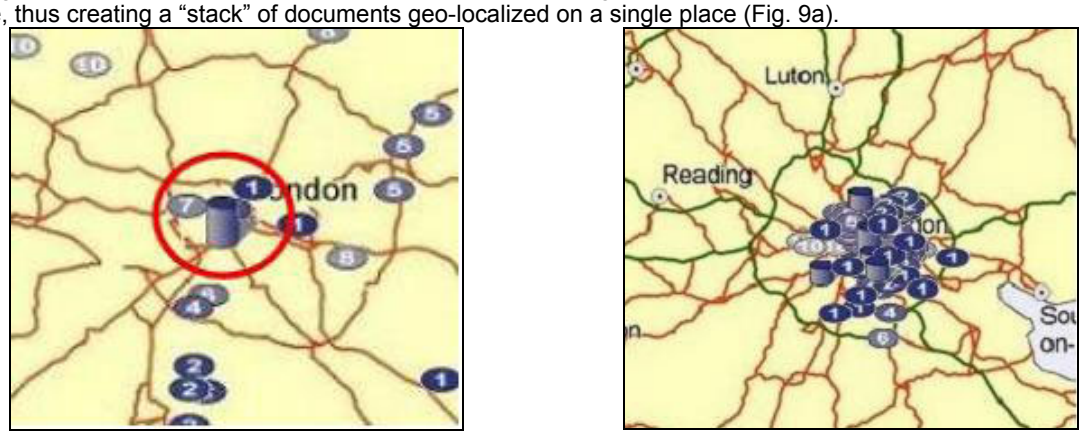

Figure 9: SPIRIT – (a) Stack of geo-localized documents – (b) Conglomerate of geo-localized documents

The extreme case relates to the space queries strongly targeted on a single place. For example, a user interested in the documents describing the town of Pau will see a stack of documents geo-localized on Pau.

For document stacks, the results remain geo-localizable and thus may be represented on maps. However the navigation modes within these sets of documents have still to be defined.

6.2.2 Representation of documents evoking very close places (conglomerates)

The problem of documents which are too close to each other relates to the documents which evoke distinct places but too close compared to the current scale to the map. This case is displayed with a conglomerate of icons located in a disordered way on a map (see Fig. 9b).

The documents belonging to conglomerates remain geolocalizable and thus may be displayed on a map. However, two problems occur: on the one hand, a problem of legibility and location for the user and, on the other hand, a problem of access to the documents because it may result difficult to click on some icons.

6.2.3 Representation of documents evoking several precise places or areas (multi-representation)

This problem relates to the documents with several absolute space features and/or any relative feature. The representation of these documents on a map can be considered in various ways and can thus pose interpretation problems for the user. Fig. 10 shows some cases where the localization of the document on a map is not elementary:

- a document evoking several absolute space features (a);
- a document evoking a relative space feature (b);
- a document both evoking both a relative space feature and an absolute space feature (c);
- a document evoking several relative space features (d).

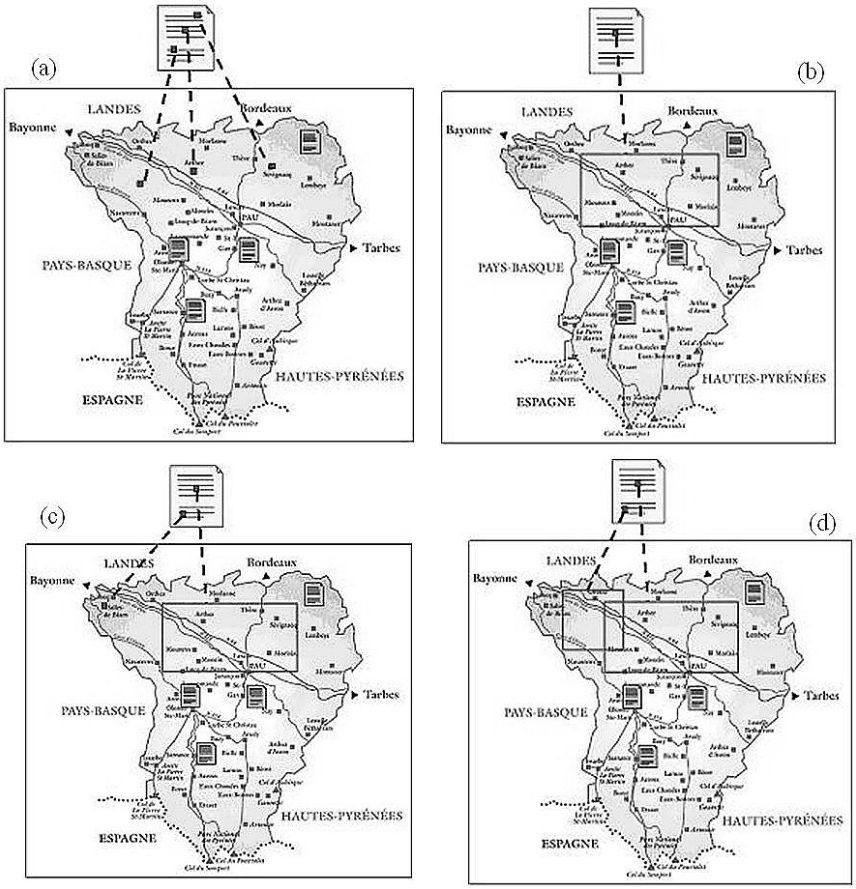

Figure 10: Multi-sites regional documents: which is the «good » representation?

The SPIRIT system uses two representation modes of these results to deal with this problem. Fig. 11 presents the set of resulting documents, at the same time on the map and in a list presented beside the map. With a mouse-over on one of the document titles belonging to the list, we can see the several places evoked in the document thanks to a colouring of the corresponding entities on the map.

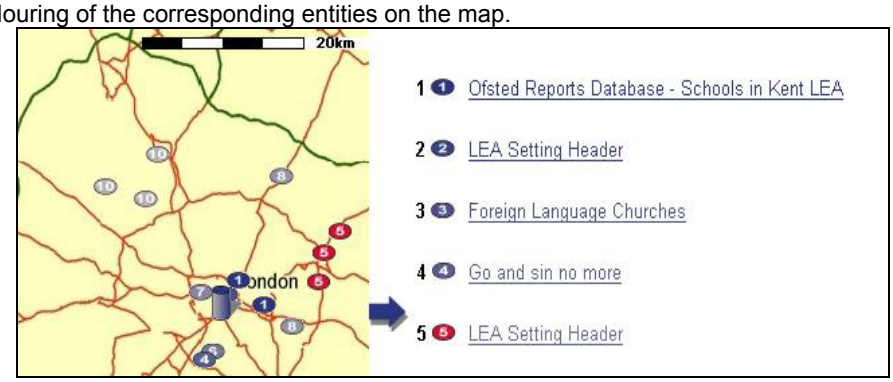

Figure 11: SPIRIT – Multi-sites documents

The problem of the documents evoking an area (rather than precise places) is not dealt with by the SPIRIT system.

6.3 Space Information Visualization within the PIV Prototype

We indicate hereafter the solutions planned to deal with these problems in the next version of the PIV prototype.

6.3.1 Conceptual Proposal

6.3.1.1 Document Stacks

In the PIV project, our approach consists in displaying any geo-localizable document on a map. In order to process all the documents belonging to a stack, we have to first identify them among all the query results and then to display them with a specific stack icon on the map. We thus keep the principle of geo-localizing the documents and we can inform the user that this place is evoked in several documents. The example presented below (Fig. 12 left side) shows a document stack (represented by a specific icon) located on the town of Pau.

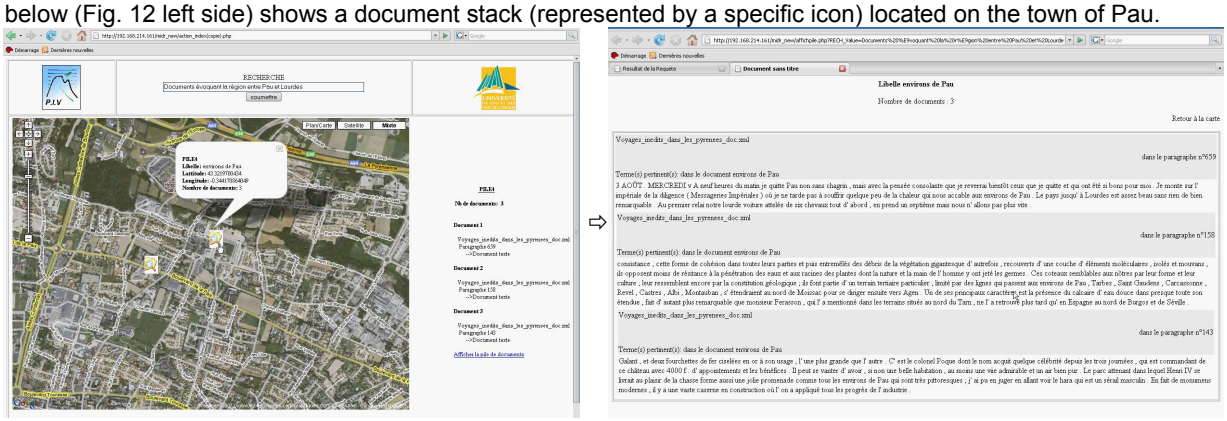

Figure 12: PIV – Stacks of geo-localized documents

The user navigation in a stack of geo-localized documents must be thought in a different way since the space discriminating criterion cannot be used to differentiate the documents. The cartographic navigation mode reaches its limits. Indeed, if the user wishes to explore a document stack (all these documents are geo-localized on a single same point), it becomes necessary to swap to another representation and navigation form: by list (Fig. 12 right side), organized by topics, etc.

6.3.1.2 Document Conglomerates

In the PIV project a document conglomerate may be found when, according to the map scale, the documents corresponding to the query results are located in almost the same area.

It thus differs from the case of document stacks and the navigation through the document conglomerate is very difficult because the screen (in pixels) distance between the corresponding icons is not sufficient to easily distinguish and access each document icon. This problem may be solved with a zoom on the concerned area. In the PIV project (Fig. 14), we are currently using the same principle as in the SPIRIT project. (Fig. 13).

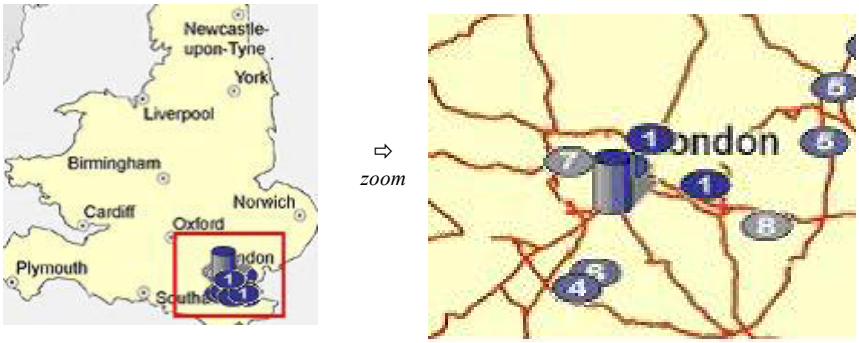

Figure 13: SPIRIT - Navigation through document conglomerate

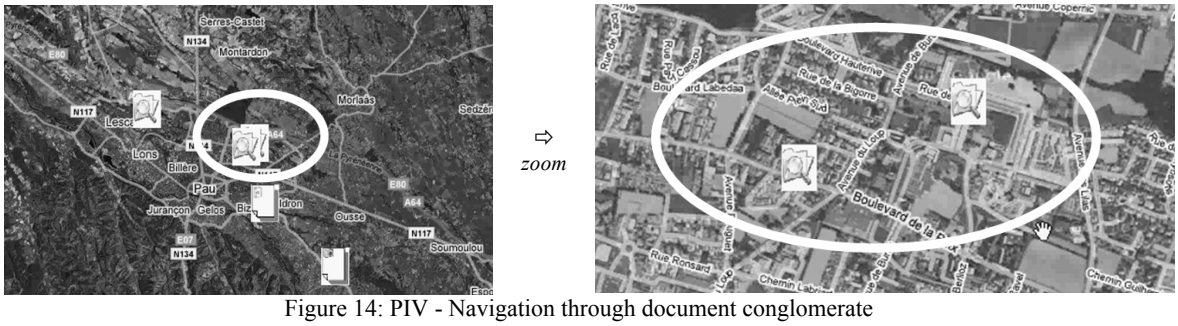

Figure 14: PIV - Navigation through document conglomerate

6.3.1.3 Multi-sites and regional documents

In the PIV project, we use the following principles:

- a document evoking a relative space feature (region) will be displayed in the centre of this region. A mouse-over on this (document) icon will display the boundaries of this region.
- a document evoking several (n) space features (either absolute or relative) will be displayed several (n) times on the map.
- all the instances of a same document will be highlighted on a mouse-over on each (document) icon to inform the user that all these icons are related to the same document.

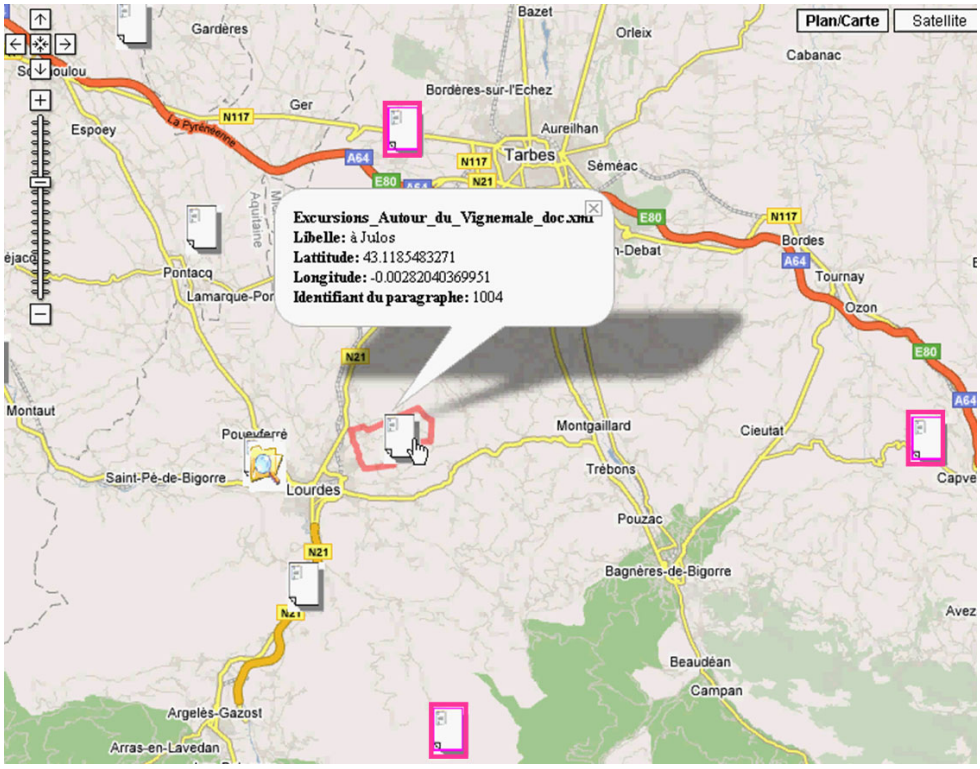

Figure 15: PIV – Regional and multi-sites documents

In Fig. 15, the user uses a mouse-over on a document icon located on the "Julos" village. The system automatically displays (light red colour) the geographical boundaries of the village and an information-bubble with the name of the place, of the corresponding document and the number of its paragraph, the latitude and longitude. Moreover the same document paragraph also deals with three other places whose corresponding icons are highlighted in pink.

6.3.2 Technical Proposal

As mentioned in sections 4.3.2 and 5.3.2, the PIV prototype is built on a flexible architecture based on web services. There are two web services combined in the information visualization process.

The first web service aims at defining and displaying the most suitable map according to the area mentioned in the user query. Firstly, it computes the geographical coordinates of the centre of all the polygons corresponding to the set of areas evoked in the user query. In a second step, it computes the geographical coordinates of the most distant space features concerned with the query. Thanks to these results, a GoogleMaps function is called to display the corresponding best map.

The second web service first calculates the way of representing each resulting documents (stacks, texts, etc.) and then displays them near the evoked place. The detection of document stacks is operated using our geographical database. PostGis provides a set of operations specialized in the processing of geographical entities. Since all the documents are indexed compared to the space entities they evoke, it is possible to detect, among all the resulting documents, those which evoke the same area and which will thus constitute a document stack on the map. Thus, the PIV system calls the intersection function of PostGis in order to find, among the resulting documents, those which evoke the same space entity (same geographical co-ordinates). A document describing several places is represented by several icons, each one located close to the evoked place. On a mouse-over on such a document, the PIV system highlights the other occurrences of this document. Conglomerate management is taken into account by scaling operations which allow the user to zoom on the grouped set of documents. These scaling operations give access to a dispersed view of the area and thus facilitate the access to each document. The implementation of the scaling operations required only little work since the maps provided by GoogleMaps[21] are also provided with the operations allowing to handle them. Since the positioning of each document (on the map) is

[21] www.google.com/apis/maps/

carried out using the GoogleMaps functions, the GoogleMaps engine itself re-computes the position of each document after a scaling operation on the map.

7. PIV Evaluation First Results

This section briefly describes the results of two experimentations carried out to evaluate PIV IE and IR processes.
A first evaluation scanned the space IE process of the PIV prototype [Sallaberry & al., 2007a]. We carried out the scanning and the OCR processes with ten books of our corpora. Then we ran the PIV prototype automatic Information Extraction processes. The PIV prototype found 9835 candidate SFs in these ten books. At the same time, we annotated geographic features like in CLEF[22] campaigns, where participants mark SFs of the samples of all these books manually. Finally, we compared these hand-written annotations to PIV results in order to compute the recall and precision rates of PIV IE processes. This study focused the evaluation on space features restricted to districts (towns and villages). Average recall and precision rates respectively equal 49% and 73%. These rates are interesting for Absolute SFs and quite bad for Relative SFs.
The stages of the IE process were analyzed and we obtained the following results:

- Within the linguistico-semantic process, the noise is weak; 80% of the SFs are detected and marked. 20% of the SFs are not detected because the syntagms responsible for their introduction are not spelled correctly, either because they are not introduced by geographic markers, or because they have no initial capital letter. RSFs like "*east of Pau*" are captured, whereas more complex ones like "*the wood is located close to the second crossroad of the main street in the direction of the Baroque Church*" are not detected. This study also pointed out the necessity of rules dedicated to inclusion relations and/or other rules skipping the stop words that separate space introducers from their corresponding named entity.
- The validation process confirmed about 30% of the detected SFs. 70% of the SFs could not actually be validated since we do not have adequate space resources. In fact, we only have resources about district type dwelling places. Another set of SFs is also lost because they are badly "OCR-ised" or spelled using old French language. This stage may be improved thanks to the extension of the resources to other SFs categories. A lexicon of old French spellings of SFs' might also improve these results.
- The ASFs geo-localization process and RSFs approximation (thanks to MBR approach) are supported by the same resources. Hence, no SF is lost here.

This experimentation led us to extend grammar rules in order to improve the RSF capturing process. We also integrated a new set of space resources describing Pyrenean roads, rivers, woods, valleys, mountains, etc.
A second evaluation presented the results of the experimentation of the PIV prototype space IR process [Sallaberry & al., 2007b]. A case study involving sample documents and queries given by the MIDR Library of Pau County made comparisons between the PIV space-based prototype and a more classic statistical-based approach.

Table 1: Space versus Classic approach results on space queries

All queries	P@5	P@10	P@15	Number of responses
A) Space approach				
Avg	0.78	0.81	0.73	637
B) Classic approach				
Avg	0.50	0.43	0.40	252

Table 2: Space versus Classic approach results on thematic and space queries

All queries	P@5	P@10	P@15	Number of responses
A) Space approach				
Avg	0.15	0.18	0.18	1 154
B) Classic approach				
Avg	0.48	0.39	0.36	331

We first evaluate the PIV prototype with space scope queries. Table 1 proves that the PIV approach brings 78% accuracy at top 5 (78% of the 5 first ranked documents are relevant) and 81% at top 10. The results showed that the PIV approach outperforms classic keywords-based approaches in the case of space queries. We also looked for the impact of using more general queries containing both space and thematic features. As it can be seen in Table 2, the results are significantly decreasing for the PIV approach (only 15% at top 5). A careful analysis of the results shows that some relevant documents are retrieved but are not ranked at the top. So, the PIV system is not suitable for rank-ordering in the case of general (space and thematic) queries. Indeed, PIV's IE and IR processes deal only with space information.
According to these results and those stated in [Vaid & al., 2005] and [Martins & al., 2005], such a space approach and statistical approaches might be combined in order to enhance retrieval accuracy in the case of general queries dealing with both space and thematic scopes. The results of an experimentation combining both approaches show a significant improvement in accuracy [Sallaberry & al., 2007b].
As the PIV system relies on an architecture based on web services, all or part of them might be easily integrated in any existing library or in any documentary management systems.

[22] www.clef-campaign.org

8. Conclusion

In this paper, we have proposed an suitable way to access documents within a corpus where the space dimension is omnipresent. There are two kinds of issues addressed in this paper. First of all, we explain how the PIV system improves information retrieval accuracy each time a query contains space criteria; second, we explain how PIV uses the space characteristics of information for result visualization.
This new access form has been designed according to the three steps involved in the design of each documentary system, that is to say: Information Extraction (IE), Information Retrieval (IR) and Information Visualization (IV).
Each part of this design framework has been tackled considering space dimension as the central criterion to favour. In order to give all its meaning to this space dimension, we have adopted a semantic approach to propose a solution built around the meaning of space. This choice has led us to set up indexing and search techniques based on the meaning of the space features found within the documents. The originality of our approach lies in the unified space model that allows one to formalize any space information whatever its expression mode (*i.e.* text, image) is. In the same way, the visual modes used to display information aim at supporting the results appropriation by using a 2D cartographic representation mode semantically close to the user's search criteria.
In order to validate our assumptions, we developed an information retrieval system operating on a corpus provided by the MIDR. Currently, this corpus is only composed of text resources (travel relations, newspapers, etc.) and of iconographic resources (lithographies). The tests carried out on this corpus deal with the evaluation of this prototype and the underlying ideas which guided its design. The first evaluation work concerned the first two modules of the data processing sequence, namely, the IE and IR steps. The results presented at chapter 7 give a first estimation of our approach interest and define the future work to undertake around the IE and IR processes. This work mainly consists in evaluating the robustness of our architecture on a bulkier corpus and composed of a larger variety of documents (postcards, sound extracts but also old movies). If this test results are not conclusive because of the too important documentary volume, we are currently studying the opportunity of integrating a step aiming to define the user expectations in a more accurate way. This stage should be performed before query processing and would aim at presenting the geographical areas concerned with the user query on a map. He/she would then have the possibility to specify (using a dedicated toolbox) the areas on which he/she wishes to obtain related documents in a more accurate way.
In addition to this work, we keep in mind the necessity to work on the issue of documents relevance, in particular about the meaning of this concept when a user specifies a query composed of space criteria.
With regard to visual aspects, we are currently working with the definition of a test protocol allowing to evaluate the relevance of the modes used to restore the results. Considering a representative population of users, the goal consists in identifying the conceptual choices which led to the implementation of problematic interactions. So, it deals with validating the methods we used to take into account the space dimension in the stage of result restitution.
This evaluation work remains necessary before considering new problems related to information visualization. Nevertheless, we keep in mind the problem of documentary navigation, in particular when a bulky corpus is considered. Currently, the navigation problem is not taken into account in our proposal; no help is provided to the user in the activity of exploring the set of resulting documents. Thus, in medium-term plans, we wish to study this browsing activity and particularly the methods to be taken into account to navigate within documents resulting from a query with space criteria. The objective of this study will aim at defining the way to take into account the space dimension to facilitate the exploration of geo-localized documents.
With the growing development of technologies such as GoogleEarth and the growing interest of users for this kind of tools, we are studying the possibility of transposing result presentation on a three-dimension map. This visualization mode would allow to increase the omnipresence of space dimension and to combine the documentary exploration with territory exploration.
Lastly, the prototype we have developed and the underlying models focus on the space dimension of documents. We keep in mind that beyond these space features, these documents, above all, contain a geographical dimension which we are not yet able to fully manage. These documents also contain temporal references associated with the descriptions of phenomena. To deal with this problem, we are currently working on a temporal model which presents some similarities with our space model. These similarities will then allow to combine the two models to index and to retrieve as well space features as temporal ones. Due to the complexity of the phenomenon dimension, we are currently working on the development of tools facilitating the manual annotation of text. These tools will be used when the tools for automatic indexing and interpretation show their limits.

Acknowledgements
The PIV project is led in partnership with the community of agglomeration of Pau and the multimedia library of Pau County (MIDR). We want to thank them for their help (by the provision of the digitized corpus, notably) and their support.
Moreover this project is shared with the whole DESI research team. That is why we want to thank its members for works performed together.

References

[Abolhassini & al., 2003] M. Abolhassani, N. Fuhr, and N., Govert, Information Extraction and Automatic Markup for XML documents, Springer, p. 159–174. 2003.

[Baeza-Yates and Ribeiro-Neto, 1999] R. A. Baeza-Yates, and B. A. Ribeiro-Neto, Modern Information Retrieval. ACM Press / Addison-Wesley, 1999.

[Benz & al., 2002] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS ready information. International Workshop - Semantic Processing of Spatial Data - Geopro, 2002.

[Bonnel and Moreau, 2005] N. Bonnel, F. Moreau, Quel avenir pour les moteurs de recherche ? In Proceedings of the workshop Manifestation des Jeunes Chercheurs francophones dans les domaines des STIC, MajecSTIC 2005, Rennes, France, pp. 291-299, november 2005.

[Bonnel & al., 2005a] N. Bonnel, A. Cotarmanac'h, and A. Morin, Gestion et visualisation des résultats d'une requête, Actes du 3e Atelier Visualisation et Extraction de Connaissances (associé à EGC'05), pp. 37-47, Paris, France, 2005.

[Bonnel & al., 2005b] N. Bonnel, A. Cotarmanac'h, and A. Morin, Meaning Metaphor for Visualizing Search Results, Proceedings of the 9th International Conference on Information Visualisation (IV'05), pp. 467-472, London, England, 2005.

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

[Boughanem & al., 2001] M. Boughanem, C. Chrisment, and M. Tmar, Mercure and MercureFiltre Applied for Web and Filtering Tasks at TREC-10. Proceeding of TREC, 2001.

[Bressan & al., 2000] S. Bressan, B.C. Ooi, and F. Lee., Global Atlas: Calibrating and Indexing Documents from the Internet in the Cartographic Paradigm. In Proceedings of the 1st International Conference on Web Information Systems Engineering, volume 1, pp. 117-124, 2000.

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

[Casenave & al., 2004] J. Casenave, C. Marquesuzaà, P. Dagorret, and M. Gaio, La revitalisation numérique du patrimoine littéraire territorialisé. In Colloque International EBSI-ENSSIB, Montréal, Octobre 2004, http://babel.enssib.fr/sommaire.php?id=612.

[Chen & al., 2006] Y-Y., Chen, T., Suel, A., Markowetz, Efficient query processing in geographic web search engines, Proceedings of the 2006 ACM SIGMOD international conference on Management of data, pp. 277 – 288, 2006.

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

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

[Cossanel & al., 2002] J. Cossanel, J.P. Cahier, M. Zacklad, and J. Charlet, Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ? In Proceedings of conférence Ingénierie des Connaissances IC2002, Rouen, Mai 2002.

[Covington, 1994] Covington, M. A. Gulp 3.1: An extension of prolog for unification-based grammar. Research report Ai-1994-06, University of Georgia Athens, 1994.

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

[Enjalbert and Gaio, 2006] P. Enjalbert, and M. Gaio, Traitements sémantiques pour l'information géographique, textes et cartes. Revue Internationale de Géomatique, to be published, 2006.

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

[Gaizauskas, 2002] R. Gaizauskas, An information extraction perspective on text mining: Tasks, technologies and prototype applications. Euromap Text Mining Seminar, Sheffield, 2002.

[Gaizauskas and Wilks, 1998] R. Gaizauskas, and Y. Wilks, Information extraction: Beyond document retrieval. Journal of Documentation, 54(1):70–105, 1998.

[Gaizauskas & al., 1995] R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks, University of sheffield: Description of the lasie system as used for muc, http://acl.ldc.upenn.edu/M/M95/M95-1017.pdf, 1995.

[Gotts and Goodday, 1996] N. Gotts, and J. Goodday, A connection based approach to common-sense topological description and reasoning. The Monist. p. 51-75. citeseer.ifi.unizh.ch/gotts95 connection.html. 1996.

[Hernandez, 2005] N. Hernandez, Ontologies pour l'aide à l'exploration d'une collection de documents. In Ingénierie des Systèmes d'Information, volume 10, pp 11–31. Hermès Sciences, 2005.

[Hill, 1999] L. Hill, Indirect geospatial referencing through place names in the digital library: Alexandria digital library experience with developing and implementing gazetteers. 62nd Annual Meeting of the American Society for Information Science, pp. 57-69. Medford, N.J.: ASIS, 1999.

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

[Ihadjadene & al., 2004a] M. Ihadjadene et collectif, Les systèmes de recherche d'informations – modèles conceptuels. Hermes, Lavoisier, ISBN 2-7462-0821-0, www.hermes-science.com, 2004.

[Ihadjadene & al., 2004b] M. Ihadjadene et collectif, Méthodes avancées pour les systèmes de recherche d'informations. Hermes, Lavoisier, ISBN 2-7462-0846-6, www.hermes-science.com, 2004.

[Jacquemin & al., 2005] C. Jacquemin, H. Folch, K. Garcia, and S. Nugier, Visualisation interactive d'espaces documentaires, Revue Information, Interaction, Intelligence, Revue en Sciences du Traitement de l'Information, Editions Cépaduès, Volume 5, n°1, 2005. http://www.revue-i3.org/volume05/numero01/ revue_i3_05_01_03.pdf

[Jones & al., 2004] C.-B. Jones, A.-I. Abdelmoty, D. Finch, G. Fu, S. Vaid, The Spirit Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. Geographic Information Science. Third International Conference, GI Science. pp. 125 – 139, 2004.

[Kules and Schneiderman, 2004] B. Kules and B. Schneiderman, Categorized graphical overviews for web search results: An exploratory study using U.S. government agencies as a meaningful and stable structure. Proceedings of the Third Annual Workshop on HCI Research in MIS, Washington, D.C., December 10-11, 2004.

[Lesbegueries & al., 2006] J. Lesbegueries, M. Gaio, P. Loustau, and C. Sallaberry, Geographical information access for non-structured data, ACM SAC Advances in Spatial and Image based Information Systems track, 2006.

[Leuski and Allan, 2000] A. Leuski and J. Allan, Lighthouse: Showing the way to relevant information, In Steven F. Roth and Daniel A. Keim, editors, *Proceedings of IEEE Symposium on Information Visualization (InfoVis'00)*, pp. 125-130, Salt Lake City, Utah, USA, October 9-10, IEEE Computer Society. 2000.

[Lowgren and Stolterman, 2004] Lowgren, J., and Stolterman E. Thoughtful Interaction Design: A Design Perspective On Information Technology, *MIT Press*, ISBN : 0262122715 (dec. 2004).

[Mackinlay & al., 1991] J. Mackinlay, G. Robertson, and S. Card, The Perspective Wall : Detail and Context Smoothly Integrated.Proc. ACM Human Factors in Computing Systems (SIGCHI'91), pp. 173-179, 1991.

[Marquesuzaà & al., 2005] C. Marquesuzaà, P. Etcheverry, and Lesbegueries, J., Exploiting geospatial markers to explore and resocialize localized documents. In Proceedings of the first International Conference on GeoSpatial Semantics, GeoS 2005, Mexico City, november 29-30, Lecture Notes in Computer Science, Vol. 3799, pp. 153-165, 2005.

[Martins & al., 2005] B. Martins, M-J. Silva, and L. Andrade, Indexing and ranking in Geo-IR systems. In Proceedings of the 2nd International Workshop on Geographic Information Retrieval (GIR), pp. 31-34, ISBN:1-59593-165-1, 2005.

[McCurley, 2001] K.S. McCurley, Geospatial Mapping and Navigation of the Web. In Proceedings of Tenth International World Wide Web Conference, Session P7. ACM press, pp. 221-229, www10.org/cdrom/papers/278/, 2001.

[Muller, 2002] P. Muller, Topological spatio-temporal reasoning and representation. Computational Intelligence, 18(3):420–450, 2002.

[Porter, 1980] M. Porter, An algorithm for suffix stripping. Program, 14(3), pp.130-137, July, 1980.

[Robertson, 1977] S.E. Robertson, The probability ranking principle in IR. Journal of Documentation 33(4), pp. 294-304, 1977.

[Robertson & al., 1998] S.E. Robertson, S. Walker, M. Hancock-Beaulieu, Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. TREC 1998, pp. 199-210, 1998

[Sagara and Kitsuregawa, 2004] T. Sagara, M. Kitsuregawa, Yellow Page driven Methods of Collecting and Scoring Spatial Web Documents. SIGIR Workshop on Geographical Information Retrieval, www.geo.unizh.ch/~rsp/gir/, 2004.

[Sallaberry & al., 2006] C. Sallaberry, C. Marquesuzaà and P. Etcheverry, Implementing a Visualization System suited to Localized Documents, Addendum Contributions to the fifth IEEE International Conference on Research, Innovation and Vision for the Future, pp. 13-18, Patrick BELLOT, Vu DUONG, Marc BUI, Bao HO Editors Editions SUGER, Collection Informatique ISBN : 2-912590-4-0, Collection Informatique ISSN : 1621-0875, Hors série ISSN : 1621-0065 RIVF, Hanoi, Vietnam, March 05-09 2007

[Sallaberry & al., 2007a] C. Sallaberry, M. Gaio, J. Lesbegueries, and P. Loustau, A Semantic Approach for Geospatial Information Extraction from Unstructured Documents. In The Geospatial Web, Springer. ISBN 1-84628-826-6. www.geospatialweb.com , 2007.

[Sallaberry & al., 2007b]. C. Sallaberry, M. Baziz, J. Lesbegueries, and M. Gaio, Towards an IE and IR system dealing with spatial information in digital libraries – Evaluation Case Study, ICEIS, 9th International Conference on Enterprise Information Systems, 2007.

[Salton and McGill, 1984] G. Salton, M.J. McGill, Introduction to modern information retrieval. McGraw-Hill Int. Book Co, 1984

[Sanderson and Kohler, 2004] M. Sanderson, and J. Kohler, Analyzing geographic queries". In Proceedings of the Workshop on Geographic Information Retrieval, SIGIR 2004, www.geo.unizh.ch/~rsp/gir/, 2004.

[Tebri, 2004] H. Tebri, Formalisation et spécification d'un système de filtrage incrémental d'information. PhD thesis, Université Paul Sabatier de Toulouse, 2004.

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

[Torres and Parkes, 2000] J.M. Torres, and A. Parkes, User modelling and adaptivity in visual information retrieval systems, Workshop on Computational Semiotics for New Media, University of Surrey, UK, June 29-30, http://www-scm.tees.ac.uk/users/p.c.fencott/ newMedia, 2000.

[Vaid & al., 2005] S. Vaid, C.-B. Jones, H. Joho, M. Sanderson, Spatio-Textual Indexing for Geographical Search on the Web. 9th International Symposium on Spatial and Temporal Databases, 2005.

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

[Wildöcher & al., 2004] A. Wildöcher, E. Faurot, F. Bilhaut, Multimodal indexation of contrastive structures in geographical documents. In RIAO, pp.555–570, 2004.

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

[Woodruff & al., 1994] A.G. Woodruff, C. Plaunt, GIPSY: Automated Geographic Indexing of Text Documents. Journal of the American Society for Information Science, 45:9:645-655, 1994.

[Zhan, 2001] F. B. Zhan, How much is region q covering region r, a little bit, somewhat, or nearly completely? M. Cristani and B. Bennett (Eds.), SVUG01: The First COSIT (Conference on Spatial Information Theory) Workshop on Spatial Vagueness, Uncertainty and Granularity, Morro Bay, CA, September 2001.

[Zipf, 1949] Zipf, Human Behaviour and the Principle of Least Effort. Addison Wesley, 1949.