Implementing a Visualization System suited to Localized Documents

Christophe Marquesuzaà, Patrick Etcheverry

Implementing a Visualization System suited to Localized Documents

Christophe Marquesuzaà
DESI Team Project
LIUPPA Laboratory
Bayonne, France
christophe.marquesuzaa@iutbayonne.univ-pau.fr

Patrick Etcheverry
DESI Team Project
LIUPPA Laboratory
Bayonne, France
patrick.etcheverry@iutbayonne.univ-pau.fr

***Abstract*—Local cultural heritage documents are characterized by contents strongly connected with a territory and its land history. Our contribution aims at enhancing such a content valorization by proposing a system allowing documents to be retrieved and visualized according to geographic criteria. The paper presents an approach for indexing and visualizing a corpus according to space criteria. After presenting the main principles of this approach, we describe the technical aspects of a prototype developed for a regional media library centre.**

Localized documents; cartographic representation; geographic information systems

I. INTRODUCTION

The problem considered in this paper deals with the design and implementation of a prototype allowing localized documents to be retrieved and visualized. In our proposal we take into account the territorial specificity of these documents to provide a smart system with a suited behavior. The work carried out deals both with indexing problems and visualization aspects. Our challenge consists in using the same approach (the space criteria) to index the documentary base, to interpret the user requests and to visualize the resulting documents. Our prototype is fully implemented and support all the designing aspects described in this paper.

The considered document management system is designed for a regional media library centre (MIDR - Médiathèque Intercommunale à Dimension Régionale). This centre aims at revitalizing its archives with a system allowing these documents to be accessed by the general public again. The richness of these documents belonging to the regional heritage cannot be shared because of their inaccessibility. These document funds generally remain in archives, museums or libraries and are used only by the few specialists who know of their existence [1].

We first present the characteristics of localized documents (paragraph 2). Then, we highlight the general characteristics of the prototype we have developed for the media library (paragraph 3). In the next two parts we focus on the method used to index the corpus (paragraph 4) and on the visualization system (paragraph 5) adopted for the results restitution. For each one of these parts, we first describe the main principles and the way to implement them from a technical point of view. We conclude this work presenting the work in progress and the future developments.

II. THE SPECIFICITY OF LOCALIZED DOCUMENTS

The documents we consider are composed of a lot of patrimonial references. The dominant property of these documents is their strong territorial attachment. The contents of text documents (tales, travel stories, etc.) teem with toponyms, place names, etc. Picture documents (postcards, etc.), most of the time, represent the characteristic places of the considered territory.

The prototype presented in the next paragraph operates with a documentary corpus of the XIXth and XXth century provided by the MIDR. Textual documents concern travel stories, tales and legends. Current graphical documents, for the moment, only consist of lithographies, maps and postcards (Fig. 1).

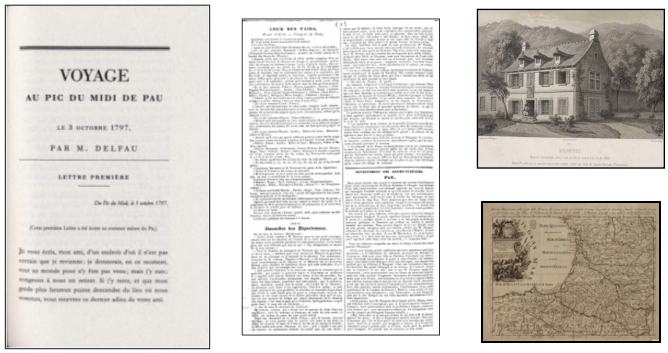

Figure 1. Some documents of the MIDR corpus

This territorial specificity, which results in the omnipresence of space within each document, allows new access forms to be imagined for document management systems. The idea consists of proposing a new access form taking into account the territorial specificity of documents. This proposal does not mean to replace traditional access forms (by author, by topic, etc.) but to enhance the current possibilities by proposing an access to the documents according to the space they evoke.

In a log study of the Excite search engine, [2] pinpoints that about one fifth of all queries contains geographical terms.

Within our specific corpus, we can say that about two fifth of our digital library queries contain geographical criteria [3].

Accessing documents according to the space they evoke aims at proposing functionalities for document research and presentation according to space criteria. It deals with the ability to consider documents evoking specific places (to seek documents evoking the town of Pau), well-known areas (to seek the documents relating to the Béarn region) or areas indicated in a more or less accurate way (to seek the documents in relation with the surroundings of Pau).

Taking into account the territorial nature of a corpus within a document management system requires (partly) an integration of the space dimension in the planned interactions with the user.

Among these interactions (Fig. 2) [4], we distinguish:

- the interactions allowing a user to express a requirement in term of information contents (Fig. 2 - A). The interactions then relate to the expression modes of the needs (text, graphics, …), the interpretation of these needs by the system, the output representation given to the user and the possible clarifying processes. These aim at bringing the real needs of the user closer to the ones interpreted by the system.
- the interactions allowing to use the information contents retrieved by the system (Fig. 2 - B). The interactions then relate to: the manner of presenting output, the possibilities offered to the user to browse/explore the restituted informative contents, the way of enhancing the location of the user during his/her navigation, and the various ways of reading/consulting the same informative contents.

Document management systems integrate the whole or part of the interactions presented above. According to their finality, some systems focus on the search functionalities and/or on the interactions aiming at making navigation easy and user-friendly.

Our contribution focuses on the visualization of territorial information in the interactions dealing with "requirements expression" and representation of results (Fig. 2 - C).

We thus study the advisability to integrate to each interaction the space dimension of the corpus the user interacts with. According to [5] new innovative tools are necessary to better access relevant information in a huge information system. Thus, the space dimension is used as a guideline to design each interaction of the PIV system presented in the next paragraph.

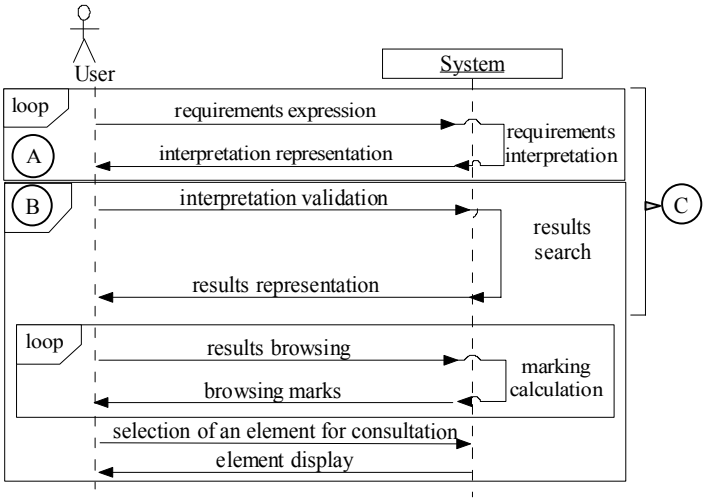

Figure 2. Interactions with a document management system

III. DISTINGUISHING FEATURES OF THE PIV SYSTEM

The PIV prototype (Pyrénées Itinéraire Virtuel –Virtual Route in the Pyrenean heritage– http://projetpiv.free.fr) is an experimental document management system which allows documents to be sought according to space criteria. In this paper, the term "space" is related with other terms used in the literacy such as "spatiality" or "spatially awareness".

The prototype is tested on a restricted corpus provided by the MIDR and composed of texts and pictures.

Starting from a textual request that expresses the user's documentary needs in terms of space criteria, the prototype carries out an analysis aiming at identifying geographical areas of interest.

A "full text" search engine –such as Google www.google.com– (Fig. 3) returns all documents explicitely containing one or several terms of the query.

Contrary and complementarily to a "full text" analysis used by most web search engines (Fig. 3), the PIV system search engine carries out a semantic analysis of the request. The PIV system is based on a semantic analysis coupled with a geographical database allowing the system to have a geographical representation of the territory. With these tools, the PIV system is able to interpret space requests in a smarter way

Figure 3. "Full text" search with Google

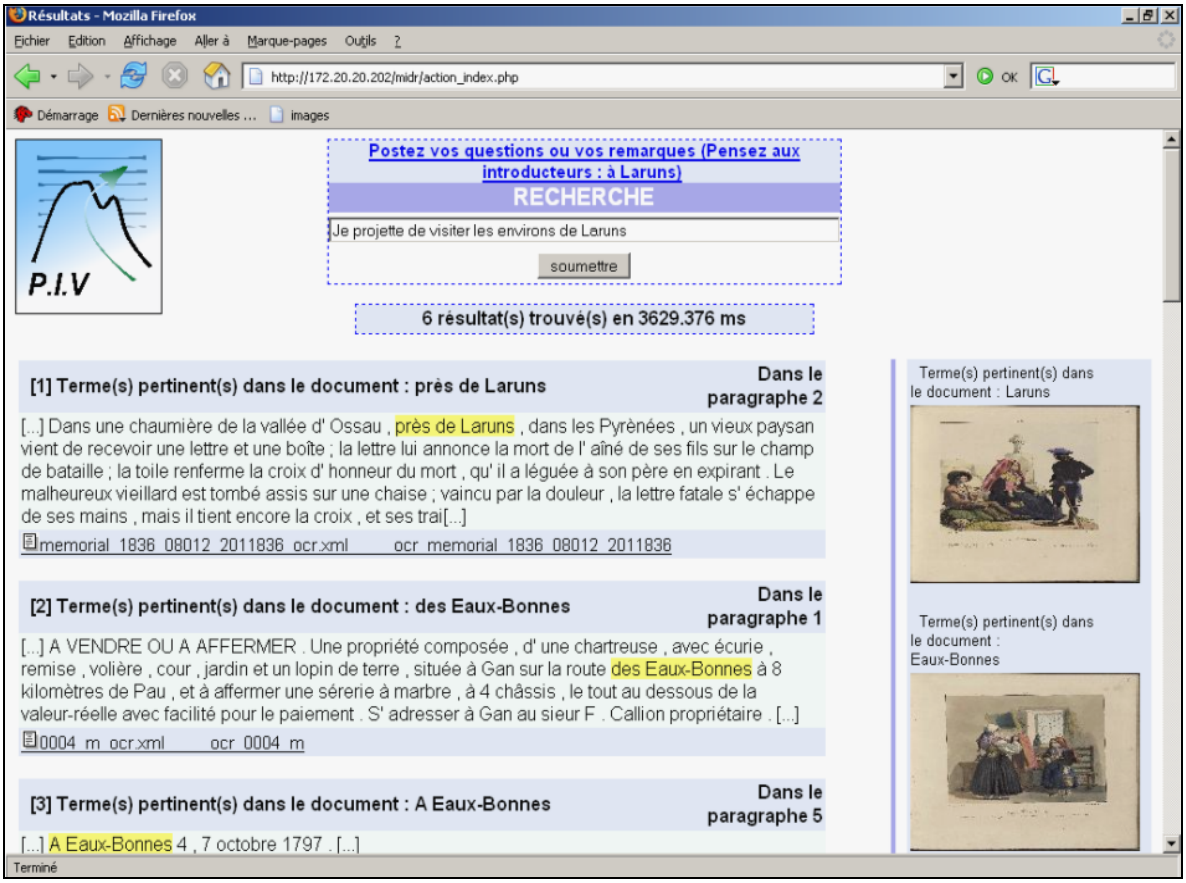

Figure 4. The PIV prototype (list visualization mode)

In the example presented on Fig. 4, the user is interested in documents evoking the surroundings of Laruns. The "*surroundings of Laruns*" are analyzed and interpreted by the system which is then able to find documents containing the terms "*close to Laruns*" ("*près de laruns*") but also place names located in the surroundings of Laruns ("*Eaux-Bonnes*").

In order to better integrate the space dimension in the stage of result restitution, we have compared the four most used approaches of information representation: list representation, topic representation, graph or tree representation and cartographic representation [6].

According to [7] "*the map is not the territory*". This is a related expression meaning that an abstraction derived from something is not the thing itself. Indeed, a map is the image of a mental representation that constitutes a link between a territory and its representation. Considering these assumptions, by opposition, [8] estimates that "*the map is the territory*". We do not want to engage into this polemic.

We think that the cartographic representation is a good way to acces the spatial information contained in our corpus. Thus, the PIV system allows a cartographic mode to be used, similarly to the SPIRIT system ("*Spatially-Aware Information Retrieval on the Internet*"–www.geo-spirit.org).

In our cartographic representation (Fig. 5), each resulting document is represented on a map near the place it evokes

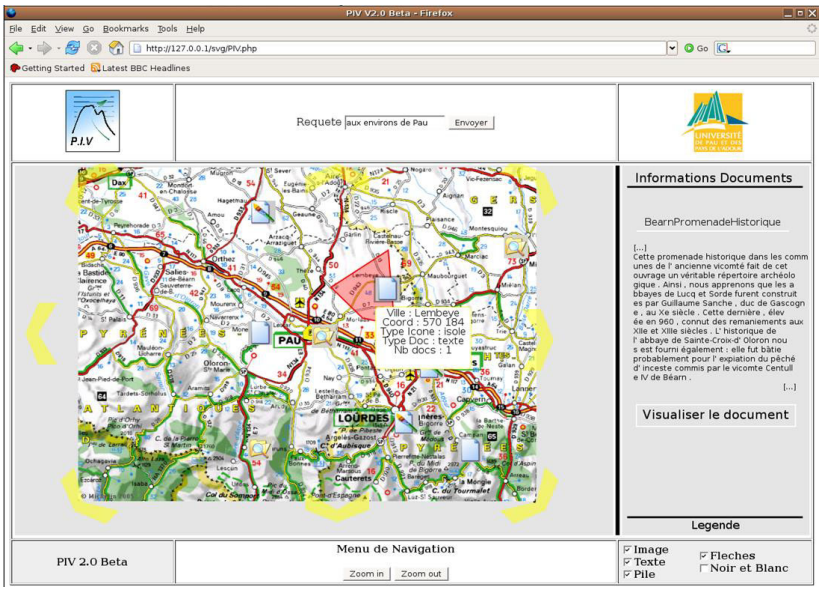

Figure 5. The PIV prototype (cartographic visualization mode)

IV. INDEXING A CORPUS ACCORDING TO SPACE CRITERIA

A. *Principles*

Finding and visualizing documents according to space criteria supposes to first index the corpus according to these criteria [9], [10]. The goal of indexing is to transform documents in substitutes able to represent the contents of these documents [11]. The indexing operation aims at locating, marking and describing (within an index) the space entities mentioned in each document of the corpus.

To complete this work, we have proposed a space model [3] describing every space feature (SF in Fig. 6 top right corner) which can be either:

- an Absolute Space Feature (A_SF) if it only consists in a named entity allowing a geo-localization, or
- a Relative Space Feature (R_SF) if it is defined using a topological relation with at least one space feature. Topological relations can be adjacency, inclusion, distance, geometric and orientation [12], [13].

This allows to describe absolute space entities ("*Laruns*") as well as relative space entities ("*close to Laruns*").

The indexing process (Fig. 6) is then carried out automatically by a set of grammatical rules [14] which seek text patterns that may refer to a space entity. Any potential space entity is then submitted to a geographical information system (GIS) which will indicate if the considered entity belongs to its database. Any selected space entity is then described within the index in accordance with the space model.

Currently, the indexing of the image type documents is carried out on this above principle. Each image has both an attached title and a library "card " filled out by the MIDR. The indexing process is then applied, not to the image but to the associated textual metadata.

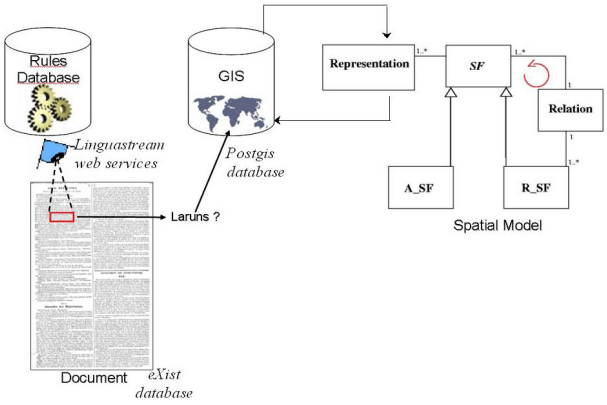

Figure 6. Space indexing of the corpus

B. *Technical Architecture for the Indexing Process*

The media library documents are stored in an eXist (http://exist.sourceforge.net) database using the xml format. The space entities identification evoked in the documents is carried out using web services based on the Linguastream (www.linguastream.org) tool.

The validation of the space entities identified by the Linguastream web services are carried out using the PostGIS (www.postgis.org) geographical information system. This database currently contains a geographical description (Lambert II étendu format) of areas and places of the French territory.

Each validated space entity is described according to our spatial model and these descriptions are stored in the PostGIS database that allows the use of geographical operators to retrieve documents according to geographical criteria: documents evoking the places located 10 kilometers from the current position, documents quoting the villages located in a given area, etc.

Table I briefly describes the informational structure used in PostGIS. The internal representation of each geographical entity corresponds to a set of points representing a polygon, which contribute to define a geographical place in an accurate way.

TABLE I. PIV INFORMATION SYSTEM FOR THE RESULTS MANAGEMENT

Element	Meaning
doc_id	Document with relevant space features for the user.
para_id	Paragraph of *doc_id* with relevant space features for the user.
sf_id	Space feature of *para_id* considered as relevant for the user
wording	Wording of the space entity. E.g. "*Laruns*" "*close to Laruns*"
sf_type	Type of *sf_id*. E.g. "*village*", "*road*", etc.
is_RSF	Space feature nature: Absolute or Relative Space Feature
the_geom	Set of points representing the polygon describing the space feature of *sf_id*.

Information treatment is not limited to documents retrieval. Once the documents retrieved, it is necessary to integrate them in a workspace. Within this framework, we need to focus on the way to restitute these results to the user in order to favour their appropriation. The visualization mode must be adapted to the type of the handled information [15], [16].

V. VISUALIZING A CORPUS ACCORDING TO SPACE CRITERIA

A. Principles

Problem of processing large amounts of text can be solved if text is spatialized in manner that takes advantage of human perceptual abilities [17].

Within our system, the user expresses his/her interest through a request expressed in natural language (1). This interest is then spatially translated in the form of geographical co-ordinates corresponding to a bounding box including the concerned area (2). Thanks to the index (3) which contains all the geographical co-ordinates of the space entities evoked in each document of the corpus (cf. section 4), the system is then able to find the resulting documents. This operation is carried out by determining which geographical entities (documents of the index (3)) have geographical co-ordinates which are intersected with the geographical co-ordinates indicating the user's zone of interest (4).

Fig. 7 summarizes the mechanism of the PIV system. The presentation of the documents in a cartographic form is carried out in two main steps:

From the co-ordinates which indicate the user's zone of interest (2) the system determines the best map representing the concerned area (5). In this phase, the system invokes a "map server" able to restore a cartographic image starting from geographical co-ordinates.

From the co-ordinates indicating the geographical entities evoked in each resulting document (3), it becomes possible to place on the map (5) each document according to the place(s) that it evokes. In this phase, the system uses a tool able to calculate, starting from the "ground co-ordinates" of an entity (extracted from the index), the "pixels co-ordinates" of this entity on the geographical map obtained at this stage (5).

From the documents presented on the map, the user can begin his/her result exploration.

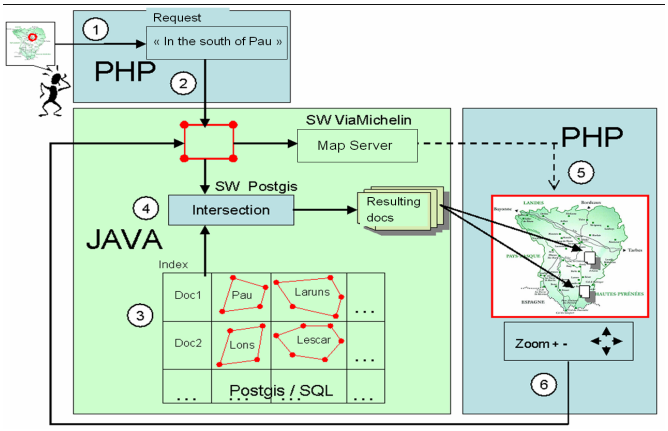

Figure 7. Overview of the PIV system

According to [18], we aim at designing and developing a system where interactions are adapted to the user needs. In the future, when the user activates a zoom or chooses to move on the map (6), this operation will consist in modifying the current geographical interest zone (2) by an arithmetic operation carried out on the current zone and taking into account the requested operation (a shift of the map to the west, etc). Once modified the current zone (2 the same process needs to be repeated as previously.

B. Technical Architecture for the Visualization Process

The PIV prototype is built on a flexible architecture based on web services in order to facilitate the integration of new modules and so, the evolution of the application.

Fig. 7 also presents the technologies used for the "front office" of the PIV system, in accordance with the previously stated principles. The technical architecture includes a module which calculates the information to be visualized and a part which recovers this information to display it.

The calculation of information allowing visualization is implemented as a web service using the Java, Apache Tomcat and Axis (http://ws.apache.org/axis) technologies.

Our web services aim at calculating information necessary to represent each request result within a web browser. The visualization of the results on a map is then carried out using PHP language and SVG (www.w3.org/TR/SVG) technology to manage the graphic aspects. The exchange of information between the Java and the PHP environment is ensured by the interoperability property of the web services.

The documents presentation on a map close to the places they evoke requires a map server able to display maps according to the corresponding geographical co-ordinates. The geolocalized presentation of the documents, according to the spaces they evoke, is carried out using ViaMichelin (http://ws.viamichelin.com) web services. These services both allow to generate the maps associated with the evoked places in the user request, and to compute the screen positions of each space entity and so, to position icons associated to each document.

VI. CONCLUSION

With a strongly localized documents collection -including lots of space references - we focus on an approach which is complementary to traditional "full text" search engines. Our contribution consists in similarly integrating geographic semantics into the next three steps: Information Extraction, Information Retrieval and Information Visualization.

Our approach allows to better integrate the user requests looking for a territory thanks to with a documentary base containing numerous space references. We have developed a fully executable prototype which completely illustrates the proposed design principles. This prototype is built on an open and flexible web services based architecture to facilitate the evolution of the application.

We will look forward on works dealing with documentary visualization [19], [20], [21]. Documentary visualization is at the crossroad of information visualization and information retrieval. From the first field, it both uses the techniques of spatial disposal of information (networks, tables, metaphors of real spaces…) and the techniques of interaction and navigation (zoom and side displacement, distortion according to the point of interest, masking and transparency, selection and filtering…). From the information research field, it uses methods of research, segmentation and classification of information.

Within our project, work still remains not only to strengthen the existing prototype by including into the documentary base much more text resources, but also by adding new handling tools put at the users' disposal. According to [22], "*users require more effective multi-dimensional visualisation tools and faster interactive performance. This challenge demands improved fundamental methods for data model and visual exploration analysis*."

Another work to perform consists in testing the prototype with heterogeneous (texts and images) documents collections and to formally evaluate the relevance of geographic query results. A first IE process evaluation [23] has just been finished and an IR process evaluation is on the go. It will also be important to integrate current work dealing with time dimension which is very present into our localized corpus. Within this scope, we will take into account work presented in [24] in order to integrate time dimension in our design method.

REFERENCES

[1] J. Casenave, C. Marquesuzaà, P. Dagorret, and M. Gaio, "La revitalisation numérique du patrimoine littéraire territorialisé." International EBSI-ENSSIB colloquium, Oct. 2004. Available: www.ebsi.umontreal.ca/rech/ebsi-enssib/ebsi-enssib-programme.html

[2] 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.

[3] J. Lesbegueries, M. Gaio, P. Loustau, and C. Sallaberry, "Geographical information access for non-structured data", 21st ACM Symposium on Applied Computing - Advances in Spatial and Image based Information Systems track, pp. 83-89, ISBN : 1-59593-108-2, 2006.

[4] P. Etcheverry, C. Marquesuzaà, and S. Corbineau, "Designing Suited Interactions for a Document Management System handling Localized Documents", SIGDOC 06, 24th ACM International Conference on Design of Communication Conference, pp. 188-195, ISBN : 1-59593-523-1, Myrtle Beach, USA, 2006.

[5] P. Brusilovsky, and C. Tasso, Preface to special issue on user modeling for Web information retrieval. "User Modeling and User Adapted Interaction", 14 (2-3), 147-157, 2004.

[6] C. Sallaberry, C. Marquesuzaa, and P. Etcheverry, "Spatial Information Management within Digital Libraries", ICDIM 06, First IEEE International Conference on Digital Information Management, pp. 465-475, ISBN : 1-4244-0682-X, Bangalore, India, 2006.

[7] A. H. S. Korzybski, "Science and Sanity, an Introduction to Non-Aristotelian Systems and General Semantics", Preface by Robert P. Pula, Institute of General Semantics, hardcover, 5th edition, ISBN 0-937298-01-8, 1994.

[8] A. Gras, "Le macro-système technique comme modèle de la mondialisation par la mise en forme des réseaux: le cas des transports aériens", PUF, Que sais-je, 1997.

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

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

[11] G. Salton and M.J. Mac Gill, "Introduction to Modern Information Retrieval", Mac Graw Hill Book Company, New-York, 1983.

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

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

[14] P. Loustau, "Traitements sémantiques de documents dans leur composante spatiale." Master's thesis of the university of Pau, 2005.

[15] G. G. Robertson, S. K. Card, and J. D. Mackinlay, "Information visualization using 3D interactive animation", Communications of the ACM, Vol. 36, Special issue on GUI, pp. 57-71, 1993.

[16] 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.

[17] J.A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, V. Crow, "Visualizing the non-visual: spatial analysis and interaction with information from text documents," Infovis, pp. 51-58, Proceedings on Information Visualization, 1995.

[18] B. Shneiderman, "Response time and display rate", in "Designing the user interface: strategies for effective human-computer interaction", Addison-Wesley, 1987.

[19] C. Jacquemin, H. Folch, K. Garcia, and S. Nugier, "Visualisation interactive d'espaces documentaires", Revue Information - Interaction - Intelligence, Vol. 5 (1), 2005.

[20] M. A Hearst, "Unsupervised non-hierarchical entropy-based clustering", In Ricardo Baeza-Yates and Berthier Ribeiro-Neto Eds, User Interfaces and Visualization, pp. 257–323. ACM Press/Addison-Wesley, 1999.

[21] M. Hascoet and M. Beaudouin-Lafon, "Visualisation interactive d'information", Revue I3, 1 :650–659, 2003.

[22] M. Jern, "Research Advances in Geovisualization and Remaining Challenges," 9th Int. Conf. on Information Visualisation (IV'05), p. 57, 2005.

[23] C. Sallaberry, M. Gaio, J. Lesbegueries and P. Loustau, "PIV: a Geographic Content-Based Documents Management System", Laboratory internal report, 2006.

[24] C. Parent, S. Spaccapietra, E., Zimányi, "Conceptual Modeling for Traditional and Spatio-Temporal Applications - The MADS Approach", Springer Eds, ISBN: 3-540-30153-4, 2006.