1

Access by Geographic Content to Textual Corpora: What Orientations?

1.1. Introduction

The volume of digital corpora is always on the rise and the retrieval of relevant documents is an increasingly delicate task. The ambiguity of natural language terms contributes to this difficulty in the automatic interpretation of the expression of the need for information as well as in the automatic evaluation of the correspondence between documents and needs. The multiple meanings of the terms and their numerous uses in varied contexts make delicate, indeed, the task of information retrieval. Our working hypothesis therefore consists of distinguishing the spatial, temporal and thematic dimensions in order to implement dedicated approaches in the processes of indexing and information retrieval (IR). The objective is to contribute to a better content analysis of textual corpora as well as to a better grasp of the search criteria expressed in a query. Let us recall that we are studying textual corpora with "territorial" denotations, digitized, to which processes of character recognition have been applied but whose logical structure has not been conserved.

This chapter is organized as follows. Section 1.2 presents the general context related to geographic information retrieval (GIR). Section 1.3 introduces privileged fields of research as well as the position of our study. Section 1.4 gives a rough sketch of our research approach in the construction of spatial, temporal and multicriteria search engines.

1.2. Access by geographic content to textual corpora

The study concerning the processing of information in text is mainly detailed in theses [BAZ 05, LES 07, PAL 10a, KER 11]. Following a number of reminders related to document retrieval and textual corpora, we will describe the characteristics of corpora with “territorial” denotations and their uses. This category of corpora will constitute the field of experimentation for our propositions.

1.2.1. *Document retrieval and textual corpora*

Document retrieval or *information retrieval* [BAE 99, BOU 08] is traditionally defined as a set of techniques allowing us to select, from a collection of documents, information that is likely to meet the needs of the user.

A *collection of documents* (document repository or corpus) is the information accessible via the document retrieval system (or information retrieval system, IRS). It consists of *documents*, unit elements. *Textual documents* are represented by a set of descriptors (terms, for example) stored in files of descriptive instructions (metadata) or indexes whose structure can be more complex [BES 04]. However, the notion of *document* in itself is vague. Generally defined by its container (e.g. a book, the physical object that contains the text), it often varies and the expected result of a query may not be an entire book but one or more particularly relevant fragments. This is indeed the reason why we use the expression “document unit” or “document fragment” to define the unit of text returned to the user [BAZ 05].

Finally, a *query* corresponds to the expression of the information needs of the user. It constitutes the input parameter to the retrieval system and is expressed in a query language that is often simple: a choice of keywords and logical operators, for instance. Nevertheless, other languages are presented in literature: natural language, graphical language, etc. [GOK 09].

1.2.2. *Textual corpora with “territorial” denotations*

A textual corpus with “territorial” denotations is composed of travelogues, stories, newspapers, novels, poems, etc. These documents describe/discuss a territory. As detailed in [KER 11], the territorial dimension is symbolized in textual documents by a significant frequency of toponyms, outlined facts or described observations. Toponyms denote, for example, streams, cities and buildings. The facts describe, for example, political or sport-related events as well as various other events. The observations refer to architecture, botany, geology, agriculture, etc. These categories of information are, in a general way, linked to a location or a period of time.

– *Territory:* The *Longman* Dictionary defines the term territory as “an area for which one person or branch of an organization is responsible”. Kergosien [KER 11] presents a consistent overview of the notion of territory. Among the different definitions proposed, we will retain the following two [KER 11, p. 70]: “A globally accepted definition in geography describes territory as a space on which an authority is exercised and is limited by political and administrative borders. This definition is subject to debate, however the notion of territory generally integrates a geographic space composed of places (spatial component) as well as relations with different subjects (thematic component) and/or references to a period (temporal component)". It also describes a second point of view, that of geomatics. "Geomatics is the scientific field hovering between geography and computer science which mainly deals with problems of storage, processing and diffusion of geographic information. The characterization of geographic information in a particular territory is defined in the form of geographic entities (GEs) composed of spatial (SEs), temporal (TEs) and thematic entities. It should be noted that each one of these entities is not always specified or can be implicit". Kergosien [KER 11] proposes an approach of ontology construction as a tool for the structured representation of a territory but also as a support to IR and to the browsing of document repositories.

– *Examples of corpora:* Territory is at the heart of numerous types of corpora. We can quote, for example, the French-speaking corpus of archives, mainly composed of texts, maps and lithographies related to the city of Saint-Étienne and to its river Furan[1]; the multi-lingual corpus (German and French) of the Swiss Alpine Club[2], composed of reports, accounts, essays and thoughts under the theme of mountaineering; tourist guides such as the different ranges of *Lonely Planet*[3] books or of the *Michelin guide*[4]; and the equally numerous hiking guides[5] and other travel blogs[6].

These corpora have the principal characteristic of containing a very large number of place names (spatial named entities will be defined further on); the places referred to in such a way generally have a fine level of detail in a relatively confined space (a river, a city and mountain range, for example). The *Geotopia*[7] and *Text+Berg digital*[8] projects are good examples of this. The objective of the first is to experiment with georeferencing techniques in order to help organize, transmit, share and interpret archival data [JOL 11]. The second aims to digitize and promote a corpus of alpine literature [VOL 10].

– *The corpus of MIDR:* MIDR[9], from a perspective of cultural heritage promotion, has digitized and implemented the optical character recognition of its heritage document repository with the aim of indexing it into a document retrieval system. This way, the digitized documents can benefit from a renewed visibility and be exploited

[1] http://umrisig.wordpress.com/les-projets/projet-geotopia.

[2] http://www.textberg.ch.

[3] http://shop.lonelyplanet.com.

[4] http://voyage.viamichelin.fr.

[5] http://www.ffrandonnee.fr/boutique/le-catalogue-des-topo-guides.aspx.

[6] http://www.blogs-de-voyage.fr.

[7] http://umrisig.wordpress.com/les-projets/projet-geotopia.

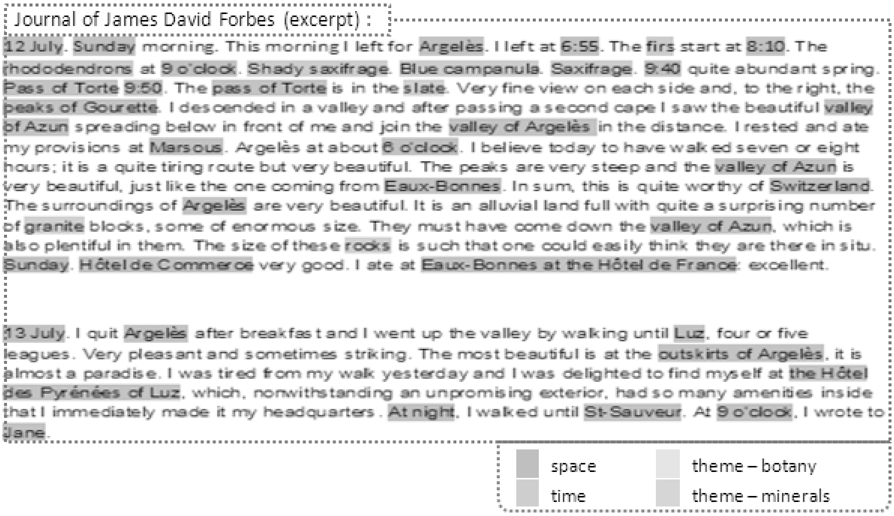
[8] http://www.textberg.ch.

[9] Médiathèque Intercommunale à Dimension Régionale de Pau Pyrénées – http://www.agglo-pau.fr.

by a larger public. It should be noted that this digitalization, keeping in mind the cost of the operation, has been carried out by a provider without the correction of errors and the recovery of the documents' structure, with the exception of their division into paragraphs.

Let us recall that this corpus is composed of documents of different types (literary studies, travelogues, newspapers, old geographic maps, lithographies, postcards, etc.), which have the common denominator of dealing with the Pyrénées territory in the 18th and 19th Centuries. A preliminary study of the corpus has revealed a predominant geographic connotation in the documents, as much in the literary studies dealing with travelogues as in the local periodicals whose articles relate to information about the territory. An experimentation has allowed us, for example, to extract almost 10,000 spatial named entities from 10 books within the corpus (i.e. 600,000 terms).

Indeed, a large amount of information makes reference to places, spatial indications as well as descriptions of landscape, temporal indicators and dates, implying a significant importance of these documents for the geographic aspect. Let us consider, as an example, travelogues (see the excerpt in Figure 1.1). The authors of these pieces of study use, most of the time, an identical structure: the text is divided into sections describing a portion of their travel. Each portion can consist of the description of an itinerary, a stage, a point of view, an observation, an event, etc.



Journal of James David Forbes (excerpt) :

12 July. Sunday morning. This morning I left for Argelès. I left at 6:55. The firs start at 8:10. The rhododendrons at 9 o'clock. Shady saxifrage. Blue campanula. Saxifrage. 9:40 quite abundant spring. Pass of Torte 9:50. The pass of Torte is in the slate. Very fine view on each side and, to the right, the peaks of Gourette. I descended in a valley and after passing a second cape I saw the beautiful valley of Azun spreading below in front of me and join the valley of Argelès in the distance. I rested and ate my provisions at Marsous. Argelès at about 6 o'clock. I believe today to have walked seven or eight hours; it is a quite tiring route but very beautiful. The peaks are very steep and the valley of Azun is very beautiful, just like the one coming from Eaux-Bonnes. In sum, this is quite worthy of Switzerland. The surroundings of Argelès are very beautiful. It is an alluvial land full with quite a surprising number of granite blocks, some of enormous size. They must have come down the valley of Azun, which is also plentiful in them. The size of these rocks is such that one could easily think they are there in situ. Sunday. Hôtel de Commerce very good. I ate at Eaux-Bonnes at the Hôtel de France: excellent.

13 July. I quit Argelès after breakfast and I went up the valley by walking until Luz, four or five leagues. Very pleasant and sometimes striking. The most beautiful is at the outskirts of Argelès, it is almost a paradise. I was tired from my walk yesterday and I was delighted to find myself at the Hôtel des Pyrénées of Luz, which, nonwithstanding an unpromising exterior, had so many amenities inside that I immediately made it my headquarters. At night, I walked until St-Sauveur. At 9 o'clock, I wrote to Jane.

Figure 1.1. *Document excerpt – The Travel to the Pyrénées, David James Forbes, CAIRN Editions (1835)*

Figure 1.1 represents two paragraphs from the travel journal of James David Forbes. In it we can find toponyms such as, for example, the "pass of Torre" whose toponymical reference name is "Torre". We can also observe temporal references such as “12 July” and thematic such as “granite” (to be considered, for example, from a mineral point of view) or “firs” (of particular interest from a botanical point of view). Let us also note the varying levels of the complexity of information: we refer to “Argelès” as a simple spatial information whereas “outskirts of Argelès” is a complex spatial piece of information that evokes a relation of adjacency whose interpretation necessitates an additional analysis of the text.

– *Geographic information:* The central element of the corpus being the geographic information, let us review a definition coming from geomatics: illustrated in Figure 1.2, it considers geographic information to be as a molecule not only composed of a spatial component, but also of a temporal component and a thematic component, or *phenomenon* [USE 96, GAI 01]. For example, the text “musical instruments in the vicinity of Laruns at the beginning of the 19th Century” fully describes this geographic molecule, with “musical instruments” corresponding to the thematic component. Let us note that some components might nonetheless be absent.

In the geography markup language (GML) specification[10] and the research on databases [LE 04], we can see the appearance of the notion of temporality. Thus, it is possible to associate a piece of geographic information with one or more geo-referenced representations, valid at a certain moment in history [GAL 01]. For instance, a city or a forest has a variable spatial definition over time which can be a creation, a disappearance, an expansion or a reduction. Finally, a phenomenon is often associated with it: subject of research, for example, a regional pollution at a given period of time [PAL 10a].

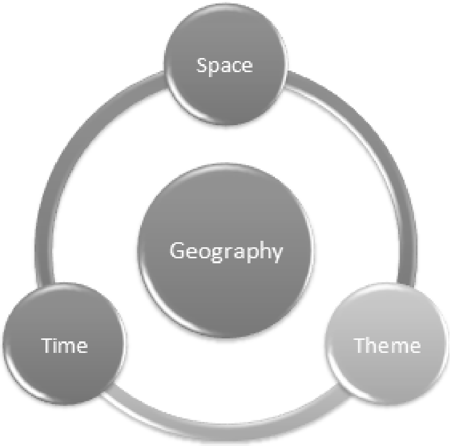


Figure 1.2. *Spatial, temporal and thematic dimensions of geographic information*

[10] http://www.opengeospatial.org/standards/gml.

1.2.3. *Access to textual content*

A study conducted on IR tasks led by students has revealed that the three main categories “of search criteria” are of bibliographical (people), chronological (periods) and spatial (toponyms) types [MAN 09]. Many other studies show a considerable proportion of references to places in the search criteria of users: for the Excite [SAN 04], AOL [GAN 08] and Yahoo [JON 08] engines, this proportion varies between 12.7 and 18.6%. Moreover, 79.5% of these queries contain toponyms [SAN 04].

In the context of digital libraries (DLs), the interfaces of IR and navigation in the resulting documents are by default composed of a subject (themes) and chronological (see the Google Books, Europeana and Gallica projects) or subject, chronological and spatial (see the Bibliothèque Numérique Mondiale project) dimensions. Here, the IR process implements advanced document management tools. These document management systems are based on metadata composed of descriptive instructions or full-text indexes in which geographical information, toponyms among others, are exploited in the same way as all the other terms.

Concerning the corpus of MIDR, a number of categories of use could be studied.

A qualitative study of the activities of librarians in the case of event-preparation scenarios has allowed us to highlight IR approaches which prioritize, in order of importance, the categories of bibliography (people), subject (themes), chronology (periods) and place (toponyms). The usage scenarios of a tourist generally prioritize the current location of the tourist or the intended place to visit in order to later focus on the subject (themes), bibliography (people) or chronology (periods). We thus distinguish three categories of users potentially involved in IR composed of geographic criteria. Their basic knowledge is, *a priori*, decreasing. The first category includes scholars, for instance historians, who wish to find precise information related to a place or a date. It also encompasses librarians, for example, whose purpose is the improvement of document annotation or the preparation of exhibitions. The second category includes the inhabitants of a region who wish to know more about it. It also affects teachers and their students, for example, who want to discover the itinerary described in a travelogue. Finally, the third category includes tourists, for example, who wish to determine the activities, the monuments or other points of interest accessible in a given zone (“the canyons to the south of Laruns”, “the springs around Pau”, etc.). It also involves every person wanting to find information from spatial and temporal criteria.

We have highlighted the significant presence of geographic criteria in the IR scenarios applied to Web content and DLs. Nevertheless, the usual search engines do not allow us to take into account the particularities of spatial and temporal information. Indeed, they are limited to the search terms (keywords) entered by the users in their query. For instance, if we wish to find documents related to events associated with the south of Pau, the search engine will target the terms “south” and “Pau”. However, a document referring to “Jurançon”, which is a bordering commune to that of Pau and situated to its south, should also be returned. Similarly, for temporal information, if we wish to find documents describing events related to the 19th Century, the search engine should not only return the documents which contain “19th Century” but also those which contain “1801”, “1802”, etc.

Finally, an experienced user interested in documents related to the “Pyrénées mountains but not those of Gavarnie, in the 19th Century, if possible, unrelated to ascents” must be able to depict this type of information need and navigate in the set of resulting text units (paragraphs). To satisfy such needs, the construction of precise indexes adapted to each type of information (spatial, temporal and thematic) seems necessary. The aim is thus to improve GIR by combining the results obtained from devoted spatial and temporal processes as well as from classic IR strategies, employed generally for thematic criteria.

1.3. Reinforcement of GIR by contributions from NLP, reasoning and multicriteria IR

If we consider the association for computing machinery (ACM) classification[11], our study is related to section *H3 INFORMATION STORAGE AND RETRIEVAL* and, in particular, to subsections: *H.3.1 Content Analysis and Indexing*, *H.3.3 Information Search and Retrieval* and *H.3.7 Digital Libraries*. It concerns IR and, in particular, GIR in textual document repositories.

However, as we have already shown, our field of research is distinct from classic IR on a large number of points. We are interested in stable textual document repositories (*a priori*, no update of a given document of the repository) as well as those which are homogeneous in their style of expression (such as *travelogue*, *walk itinerary* and *tourist guide*). This particularity enables a thorough processing, on the one hand, for *back-office* indexing and specific usage scenarios, on the other. Concerning indexing, natural language processing (NLP) supports the targeted extraction and analysis of spatial and temporal information, while qualitative reasoning completes this analysis and supports the interpretation of this information as well as that of associated relations. Thus, Figure 1.3 positions our study concerning GIR at a cross-road between IR, NLP and qualitative reasoning. This can involve specialized IR dedicated to vocabulary proper to the expression of space and time. We propose an active parsing of the textual document, in other words a targeted search of expected elements of information in the text in order to build the corresponding spatial and temporal meaning of the speech.

[11] ACM Computing Classification System – http://dl.acm.org.

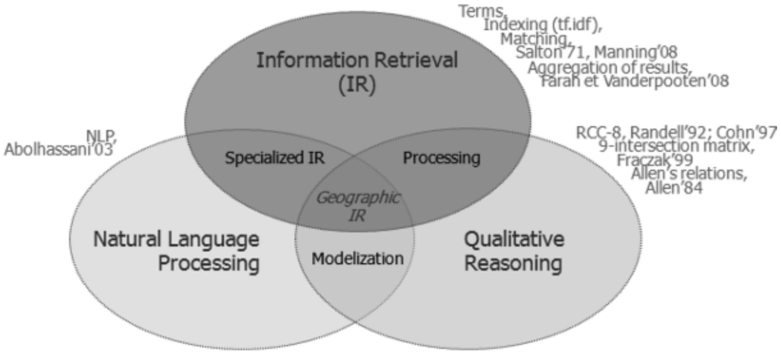


Figure 1.3. *Our GIR study and the main research fields of interest involved – extended extract of Julien Lesbegueries [LES 07]*

In Figure 1.3, we also show the necessity of having spatial and temporal models for the description of information extracted from texts. A number of processes, supported by geographic information systems (GIS), for example, dedicated to spatial and temporal information will allow us to calculate numeric representations (geometries, calendar periods) corresponding to the information described in the models.

Before describing our approach, which combines models and processes drawn from these different fields for the construction of a multicriteria search engine, we present the notions of IR, GIR, NLP and qualitative reasoning, some of which have already been briefly introduced.

Manning *et al.* [MAN 08b] define IR in two main processes with, on the one hand, techniques of document indexing and retrieval mechanisms, on the other. Thus (see section 1.2), the documents are represented by a set of descriptors (terms, for example) organized in an index. The user's need (query) is dealt with in a similar way: it is also represented by a set of descriptors. Afterwards, matching mechanisms compare the descriptors of the query with those contained in the index in order to build a list of relevant documents.

GIR differs from IR in its explicit recognition and modelization of space and time in the frame of indexing processes and IR [LEI 07, ALO 08]. In a GIR system, not only the key terms are indexed, but also the "spatial terms" with the corresponding geo-localizations called *spatial footprints* and the "temporal terms" with the corresponding intervals of time called *time stamps*. IR is, in this case, completed by the comparison of the spatial footprint or the time stamp of a query with the spatial footprints or time stamps extracted from the documents. In general, the comparison is based on the intersection sizes of these footprints and stamps.

The recognition of "spatial" and "temporal" terms in texts is supported by techniques of named entity recognition (NER). NER, detailed in [CHI 98], consists of the retrieval of textual objects called named entities (in other words, proper nouns, expressions of time and numeric expressions) which can be categorized into classes, such as people's names, names of organizations or businesses, place names, quantities, distances, values, dates, acronyms and abbreviations.

Moreover, the studies of linguistic analysis have found applications in the world of IR: for instance, the concept of *target/site* described by Vandeloise [VAN 86] and that of *concrete entity/reference* described by Borillo [BOR 98] shows the particular way humans describe spatial information when it comes to writing. It is more and more common to see techniques of linguistic analysis being associated with techniques of statistical analysis [DEL 04]: for example, the detection of named entities in a text uses morpho-syntactical processes of linguistic analysis [MIK 99]. Thus, the tools of NLP support a fine analysis based on the interpretation of the semantics contained in the textual documents. They contribute to NER and to the extraction of noun phrases which contain these entities. For example, "the south of Pau" is the phrase evoking the spatial named entity *Pau*, and "in the beginning of the 18th Century" is the phrase evoking the temporal named entity 18th Century.

Qualitative spatial reasoning (*QSR*) and qualitative temporal reasoning (*QTR*) complete the study of language by proposing reasoning processes for the acquisition of additional knowledge. The importance given to the qualitative aspects of spatial information stems from ancient Greece, as Kowalski *et al.* [KOW 07] recall. More recently, the study carried out by Allen [ALL 91] focuses on the temporal reasoning for a qualitative representation. Propositions for QSR have then adapted this study by taking into account the specificities and the bigger complexity of spatial information. Cohn [COH 96] and then later Cohn and Hazarika [COH 01] show the state of the art of QSR and, in particular, classify the spatial relations (see section 2.3.4). In the context of GIR, a reinvestment of the study related to QSR and QTR targets the interpretation of spatial and temporal noun phrases, respectively.

1.4. Toward the construction of a multicriteria IR engine

Accessing the content of textual documents via an IR approach integrating the spatial, temporal and thematic dimensions (Figure 1.2) is the main challenge of this study. Its objective is the construction of an IR engine combining these three dimensions.

1.4.1. *Challenges, hypotheses and research objectives*

1) *Challenges:* As we mentioned previously, the ratio of queries integrating spatial criteria, for example, varies between 12.7 and 18.6% according to Excite [SAN 04], AOL [GAN 08] and Yahoo [JON 08]. Although well tried today, the classic approaches to IR are limited in the case of geographic search criteria [LIE 09]. It remains true that what interests humans most often is the theme. However, taking into account the thematic dimension and more importantly the semantic aspects it holds is a very difficult task. Indeed, the current tools of IR are efficient but limited to terms. Our first objective, therefore, is to target the spatial and temporal dimensions as privileged entry points in the texts. The aim is, thus, to complete classic IRSs by specific services dedicated to the spatial and temporal aspects. The first challenge is therefore: *What models of representation and retrieval of spatial and temporal information should be proposed for the access by geographic content to textual corpora?*

It is then a question of combining a classic IRS with spatial and temporal IRSs. The heterogeneity of the models of representation and those of the corresponding IR does not allow us to directly consider the combination of such systems. The second challenge is thus: *Which core model of representation and retrieval of information should be proposed in order to prepare for the combination?*

Finally, the combination has to be based on aggregation operators adapted to the geographic context. Let us not forget the need for a power of expression highlighted earlier. This need is partially satisfied by the operators proper to each dimension. Nonetheless, it has to be completed with a finer formulation of each criterion: is it mandatory, is it associated with a level of preference, is it a rejection criterion? The third challenge is thus: *What advanced aggregation operators should be proposed and implemented for a multicriteria IR combining many IRSs?*

2) *Hypotheses:* We issue a study hypothesis for each of the challenges evoked. Concerning the implementation of specific process flows dedicated to spatial and temporal information, our study hypotheses are the following:

– A dedicated spatial IRS gives better results than a classic thematic IRS for IR composed of only spatial criteria.

– A dedicated temporal IRS gives better results than a classic thematic IRS for IR composed of only temporal criteria.

– A “rough” coupling of spatial, temporal and thematic IRSs gives better results than a classic thematic IRS for multicriteria IR despite the numerous possible biases linked, for example, to the heterogeneity of value domains with manipulated scores.

Finally, the implementation of a multicriteria IRS combining the results from spatial, temporal and thematic IRSs arises the following study hypotheses:

– The generalization of data representation is adapted to the spatial and temporal information.

– The generalization associated with a well-tried IR model does not imply a loss of quality with respect to the initial dedicated IRSs.

– A “classic” coupling (an arithmetic mean, for example) of spatial, temporal and thematic IRSs gives better results than a single thematic IRS or a two-by-two pairing of spatial, temporal and thematic IRSs.

– An “advanced” coupling (offering greater power of expression to the user) of spatial, temporal and thematic IRSs gives better results than a “classic” coupling.

3) *Objectives:* The research objectives corresponding to the formulated challenges mainly target the design and implementation of:

– models supporting the symbolic and numeric representation of spatial and temporal data;

– process flows of spatial and temporal information extraction, interpretation and indexing;

– spatial and temporal IR models;

– a model of generalization of the representation of indexed data;

– a model of multicriteria IR.

1.4.2. *Approach*

Our approach is divided into two main steps. First (Figure 1.4), under the direction of Gaio [GAI 08], we have implemented process flows dedicated to the automatic recognition of spatial and temporal named entities. We consider that, regardless of the territory, it is possible to clutch to a single or a set of named entities as characteristic elements of the content of a corpus of documents. The processes, successively lexical, morpho-syntactical and semantic consist, respectively, of extracting, validating and interpreting spatial and temporal named entities for the purposes of annotation and/or indexing. We propose a spatial and a temporal model dedicated to the needs of marking and representing the processed information in each of these steps. We obtain, in particular, spatial and temporal indexes in compliance with these models.

Texts annotated in such a way mainly target experts: librarians or archivists, for example, who need these results in order to create descriptive instructions (metadata) in a semi-automatic way. In addition, the indexes represent the content of document repositories and support IR scenarios as well as more sophisticated approaches for the discovery of territory targeting, for example pupils, tourists or scholars. We propose dedicated models of spatial and temporal IR.

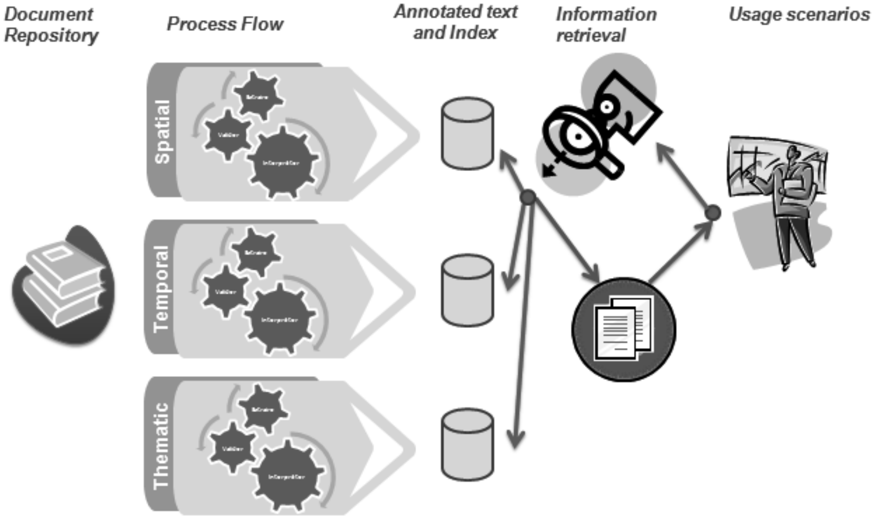


Figure 1.4. *Spatial and temporal process flows (extraction, validation and interpretation)*

Second (Figure 1.5), in continuity with the first step, we have implemented a multi-dimensional and a multicriteria IR model. We propose submitting each criterion of a query to the IRS dedicated to the corresponding spatial, temporal and thematic dimensions, followed by the combination of the results. Before any combination, however, in order to avoid possible biases, we have chosen to generalize the representation of the data corresponding to each dimension. This generalization necessitates first and foremost the segmentation of the space (respectively the period) covered by the document corpus to be indexed: we call this spatial (respectively temporal) tiling, or splitting. This results in generalized indexes (Figure 1.5). We then proceed to a *projection* in which every intersection between a tile of the generalized index and an object of the initial index increases the weight of the tile. We propose regular, administrative and calendar tilings with tiles of various size.

Thus, this tiling approach, comparable to the generalization by truncation or lemmatization of terms in classic IR approaches, allows us the implementation of well-tried IR models for each of the geographic dimensions. We compare, for example, vectorial IR to the ad hoc IR models developed for each dimension. Losses in precision and recall are of course induced by the generalization. Nevertheless, the integration of tile reference frequencies in the calculation of weight and relevance scores delivers gains that we have quantified (see section 3.5.2, Chapter 3).

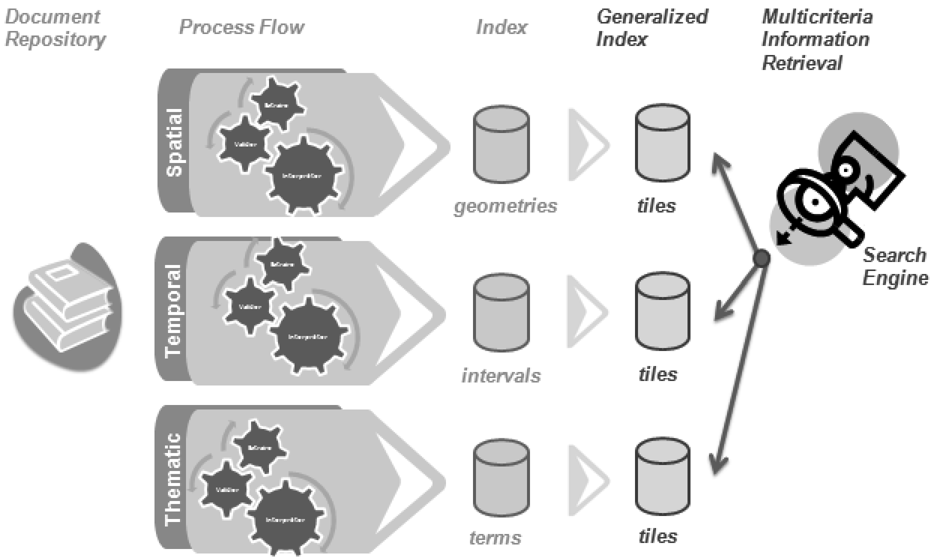


Figure 1.5. *Spatial and temporal generalization for multi-dimensional and multicriteria IR*

We now have models of representation and processing of normalized data for every geographic dimension. We design and implement a multicriteria IR meta-engine combining the results of the spatial, temporal and thematic IRSs (Figure 1.5). These are the generalized spatial and temporal IRSs. The thematic IRS is limited for now to a "basic" engine corresponding to the Terrier IRS [OUN 05].

Such a tool of multicriteria IR targets at least two categories of users. The experts benefit from new operators whose objective is to associate a stronger power of expression with each criterion of the query and thus influence the results of aggregation algorithm. Occasional users will employ predefined aggregation functions that do not need particular skills.

1.4.3. *Applications*

This study is, at the same time, motivated by and contributing to numerous projects (Figure 1.6). The GEOSEM2 project (*Interdisciplinary Program Information Society – CNRS*) has allowed us to experiment with the TAL platform LINGUASTREAM [BIL 06b] and develop spatial information annotating tools in the specific context of document repositories digitized by the Pau Pyrénées Multimedia Library (MIDR). In the continuity of these first results, the PIV project (*Pau Pyrénées agglomeration Community Project*) consists mainly of the development of two process flows dedicated, respectively, to the extraction of spatial and temporal information contained in textual documents. In the frame of the GEOTOPIA project (*CNRS TGE–ADONIS Program – Innovative tools of digital processing for the valuation and diffusion of data*), we adapt and encapsulate these processes in Web services dedicated to spatial and temporal-annotations of documents in a Web platform of sharing and collaborative enrichment of archival data. The GEONTO project (*ANR-07-MDCO-005-01 Project – Flow of Data and Knowledge*) leads us to use the "public" document repository digitized by MIDR for the purposes of enriching a geographic ontology. This geographic ontology, resulting from the processing of numerous resources, is then integrated into the spatial process flow for a better typing of the toponyms and consequently a more adapted approach of disambiguation during the search for representations in gazetteer-like geographic resources (dictionary of place names, describing their physical characteristics – nature or geo-localization of places – as well as social and economical). The GEOCIME project (*64 – General Council Project*) targets the development of an educational application allowing children and their educators to study on collections of multimedia documents before, during and after a field trip. It is based on indexing tools resulting from these different projects. The CIDRI project (*Pau Pyrénées agglomeration Community Project*) is the extension of the PIV project. It aims to generalize spatial and temporal indexes in order to develop an IR prototype combining spatial, temporal and thematic criteria (limited to terms). The MOANO project (*ANR-2010-CORD-024-01 Project – Content and Interactions*), in its GIR aspect, targets the construction and integration of a domain-specific ontology dedicated to botany. This ontology serves as a core model in the generalization of indexes of the thematic process flow. The thematic point of view on the document corpus used is here limited to botany. Finally, the IR engine prototypes PIV^1, PIV^2 and PIV^3 result from these different pieces of study. PIV^1 proposes a spatial IR engine and a temporal IR engine based, respectively, on fine spatial footprint and time stamp representations. PIV^2 proposes a spatial IR engine and a temporal IR engine based, respectively, on generalized representations corresponding to spatial and temporal tilings. Thus, PIV^2 implements the vectorial IR model in both prototypes. PIV^3 proposes a multicriteria IR engine using spatial PIV^2, temporal PIV^2 and Terrier IRS [OUN 05]. PIV^3 then aggregates the results coming from these different IRSs.

All these applications are also motivations for our approach. We can thus say, by recalling our research objectives mentioned earlier, that:

– our models support the symbolic and numeric representation of spatial and temporal data, as well as our spatio-temporal processes of extraction, interpretation and indexing are put to the test in the frame of the GEOSEM2, GEOTOPIA, GEONTO and PIV projects;

– our spatial and temporal IR models are put in practice in the frame of the PIV, GEOCIME and PIV^1 projects;

– our models of generalization and representation of indexed data as well as our models of multicriteria IR are put to experiment in the frame of the CIDRI, PIV^2, MOANO and PIV^3 projects.

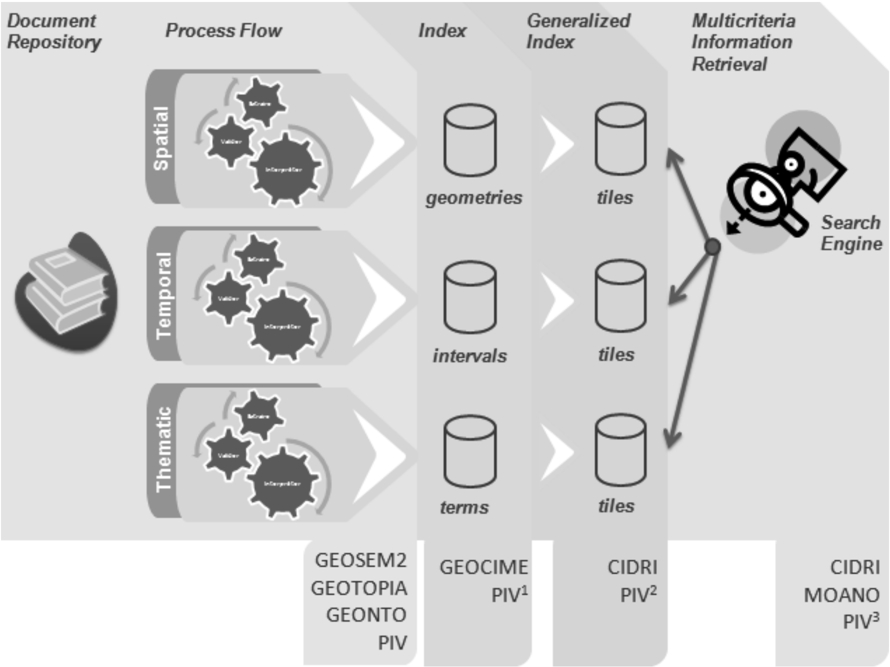


Figure 1.6. *Integration of the study into projects supported by the ANR, CNRS and local authorities*