When Textual Information Becomes Spatial Information Compatible with Satellite Images

Eric Kergosien[1], Hugo Alatrista-Salas[2], Mauro Gaio[3],
Fábio N. Güttler[4], Mathieu Roche[5] and Maguelonne Teisseire[5]
[1]*GERiiCO, Univ. Lille, Lille, France*
[2]*Pontificia Universidad Católica del Perú, Lima, Peru*
[3]*LIUPPA, Univ. Pau, Pau, France*
[4]*ICube, Univ. Strasbourg, Strasbourg, France*
[5]*UMR TETIS & LIRMM, Montpellier, France*

Keywords: Natural Language Processing, Information Retrieval, Spatial Information, Land-use Planning.

Abstract: With the amount of textual data available on the web, new methodologies of *knowledge extraction* domain are provided. Some original methods allow the users to combine different types of data in order to extract relevant information. In this context, we present the cornerstone of manipulations on textual documents and their preparation for extracting compatible spatial information with those contained in satellite images. The term *footprint* is defined and its extraction is performed. In this paper, we describe the general process and some experiments conducted in the ANIMITEX project, which aims to match the information coming from texts with those of satellite images.

1 INTRODUCTION

Nowadays, impressive amounts of high spatial resolution satellite data are available. This raises the issue of fast and effective satellite images analysis as it still requires a costly human implication. In this context, automated and semi-automated remote sensing approaches attempt to tackle this challenge.

Such large amount of data is also related to increasing temporal repetitivity of Earth Observation satellites (i.e. one image every 5 days for the new satellite Sentinel-2 instead of 16 days for the current Landsat-8). At the same time, the web offers a large amount of textual data and many researcher communities are interested in the issue of knowledge extraction including spatial information.

The ANIMITEX project aims at processing massive and heterogeneous textual data (i.e. *big data* context) in order to provide relevant information to enrich the analysis of satellite images. The project has many application areas such as image annotation (Forestier et al., 2012). For instance, identifying the precise type of crop or the function of a building is not always possible using only remote sensing images. Nevertheless, textual data could contain this kind of information and give additional meaning to the images. The development of approaches based on image/text matching becomes crucial in order to complete image analysis tasks. It also enables a better classification of data. Moreover, image-text matching will enrich Information Retrieval (IR) methods and it will provide users a more global context of data (Sallaberry et al., 2008). This can be useful for experts involved in land-use planning and management.

In this paper, we investigate one specific scenario: The construction of a bypass north of Villeveyrac (a small town close to Montpellier, south of France). The aim of this case study is to show how to enrich images with spatial information present in newspaper articles provided by Midi Libre (French newspaper). The main difficulty in extracting spatial information, e.g. Spatial Features (SF), is the ambiguity inherent in natural language. SF extraction methods generally exploit two complementary fonctionalities: toponym recognition (geoparsing) and toponym resolution (geocoding) (Leidner and Lieberman, 2011). The problem of the recognition of toponyms (place names) in text, can be seen as a particular category of named entity recognition and classification (NERC). According to Smith and Mann (Smith and Mann,

Kergosien, E., Alatrista-Salas, H., Gaio, M., Güttler, F., Roche, M. and Teisseire, M..
When Textual Information Becomes Spatial Information Compatible with Satellite Images.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 301-306
ISBN: 978-989-758-158-8

2003) there are actually several types of ambiguity involved in toponym resolution (i.e. associate a toponym with its spatial representation). The use of contextual elements (other than toponyms), such as words that have a geographical denotation (”river”, ”town”, ”basin” , etc.), can be extremely important in a toponym disambiguation task (Hollenstein and Purves, 2010).

This paper focuses on the use and adaptation of Natural Language Processing (NLP) techniques for recognition of the spatial representation throughout the document. To achieve this, we have collected a set of newspaper articles (corpus of 3809 textual documents) about the *Thau basin* territory from 2010 up to 2013. NLP methods based on lexico-syntactic patterns (Gaio and Nguyen, 2011) were then used to automatically annotate linguistic expressions conveying more or less complex spatial information. In the proposed approach, SF appearing in a text, are composed of at least one **named-entity allowing a geolocation** and one or more spatial indicators specifying its location (Lesbegueries et al., 2006). Once the SF extracted, the problem is to identify their spatial characteristics in order to define a spatial representation throughout the document.

The paper is structured as follows. In Section 2, an overview of SF extraction methods is presented. In Section 3, the method to identify SF representation is detailed. Section 4 gives a short description of the corpus, reports experiments and lists associated prospects. The paper ends with conclusions and some perspectives.

2 RELATED WORK

NERC methods automatically annotate different types of named entities: dates, people, organisations, themes, numeric values, as well as place names. There is a significant number of systems available, both proprietary and open source, such as OpenNLP[1] from Apache, OpenCalais[2] from Thomson Reuters, and CasEN (Maurel et al., 2011). More specific methods that are solely concerned with geographical data are known as geoparsing (Leidner and Lieberman, 2011). In our work, we focus on this category and a first issue is to precisely identify named-entities allowing a geolocation using the definition proposed in (Lesbegueries et al., 2006).

In this model, SF can then be identified in two different ways:

- an **Absolute Spatial Feature (*A_SF*)** one Named-Entity (NE) allowing a geolocation, such as:

  $$< (spatialIndicator)^{*},\ NE\ of\ Location >$$

  A *spatialIndicator* is a term contained within a geographic lexicon (“river”, “town”, “mountain”, etc.). Two examples of this type of SF are the ”The Thau basin” and ”the town of Sète”;

- a **Relative Spatial Feature (*R_SF*)** one spatial relationship (topological or Euclidean) with at least one SF. An *R_SF*, including one *A_SF* at the end of a pattern, is defined as:

  $$< (spatialRelation)^{1..*}, A\_SF > \quad \text{or} \quad < (spatialRelation)^{1..*}, R\_SF >.$$

  Five spatial relation types are considered: orientation (“in the south of”, etc.), distance (“20 kilometres from”, etc.), adjacency (“near”, etc.) , inclusion (“in”, etc.) , and geometry which defines the union or intersection linking two SFs (between A and B, etc.). An example of this type of SF is “in the area of Cuzco” according to the pattern $< (spatialRelation)^{1..*}, A\_SF >$.

A second issue is related to the identification of spatial representation for each SF. In this sub-domain, first research works in the 90s have been focused on the representation of complex qualitative relationships such as orientation (Frank, 1991). The direction of a SF is defined taking as reference the position of a second SF. To achieve this, the author proposes the model with cones to represent the four cardinal points: north, east, south, west. A second proposal was to represent the orientation relationship using a 3x3 grid in which the central cell is called the ”neutral zone”. Cells around represent eight cardinal points, north, northeast, east on southeast, etc. In (Hernández et al., 1995), the authors focus on the study of the representation of qualitative distances (far, near, etc.) and propose a representation model of for flexible distances at different levels of granularity. In (Cohn, 1996), a state-of-the-art of SF representation is drawn. At first, the author focuses on the use of an ontology of geographic objects and presents a study on the extension of basic physical objects (points, lines, etc..) to build more complex figures (regions , roads, etc..). In a second step, the author takes into account the topological relationships between SF and relations (orientation, distance, size and shape of related objects). These representations, enabling to take into account the underlying abstractions, are then used to create complex qualitative models.

More recently, in (Davis, 2013), the authors rely on a corpus of citations to identify the spatial relationships between spatial objects. The authors describe

[1] https://opennlp.apache.org/

[2] http://www.opencalais.com/

several problems related to the identification of relationships between SF, such as ambiguity in the recognition of the spatial representation, continuity of geometrical objects and many others.

3 SPATIAL FOOTPRINT AND DOCUMENTS

3.1 General Process to Match Satellite Images and Textual Documents

In the context of ANIMITEX project, we propose a new approach which is designed to semi-automatically match satellite images and textual documents in land-use planning contexts. This general approach is divided into three stages (see Figure 1). In the first stage, the approach focuses on the semi-automatic extraction of image objects from satellite images, using temporal clustering and segmentation. In the second stage, NLP methods have been applied in order to identify linguistic features concerning spatial information in the documents. The use of lexicons and dedicated rules (Gaio and Nguyen, 2011) allows us (1) to identify the absolute (e.g., "Madrid") and relative (e.g., "south of Madrid") Spatial Features (AS_F and *R_SF*) and (2) to calculate the footprint of each document. In the final stage, we propose to select textual documents in relation with the studied land-use planning project.

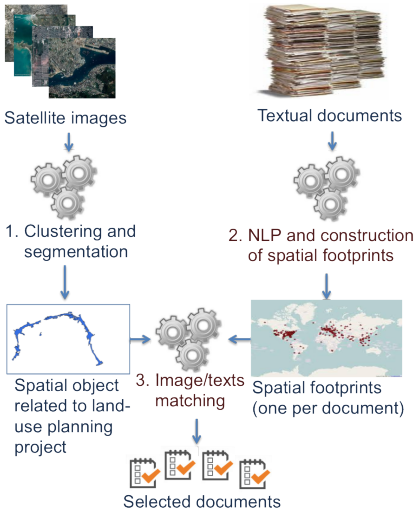

Figure 1: General process of the ANIMITEX project.

In this paper, we focus on the second and third stages. The following section describes the proposed approach to construct footprints from textual documents, and then to match these documents with satellite images using spatial representation.

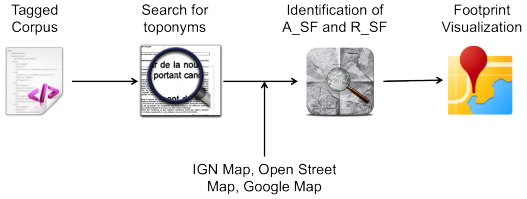

Figure 2: From document to footprint definition: the three-steps process.

3.2 Footprint Identification Process

To calculate the footprint, we propose a process consisting of three steps (see Figure 2).

SF Marking: the text marking step is achieved through an automatic framework containing all specifications proposed in (Gaio and Nguyen, 2011). This framework outputs a structure incorporating the SF definition proposed in the previous section. A list of *A_SF* and *R_SF* can be extracted from each document of the treated corpus.

SF Validation: for each identified *A_SF*, we check on external resources if there is a corresponding spatial representation. In particular, we have used layers provided by the IGN[3] (municipalities, roads, railways, buildings, etc.). In addition, if an *A_SF* does not present on IGN resources, we use gazetteers (Geonames and Open Street Maps) to complete the information. Concerning the representation of *R_SF*, we use spatial indicators of topological order associates to *A_SF*.

Following the scopes proposed in (Sallaberry et al., 2008), the spatial indicators of topological order have been grouped in five categories:

• *Proximity*: Different indicators can be used in relationship of proximity, such as "near", "around", "beside", "close to", "periphery", etc. To represent this relationship, we define a Tolerance Zone *TZ*. This area built around the studied *A_SF* will be calculated as follows:

$$TZ(A\_SF) = surface(A\_SF) + \phi * surface(A\_SF)$$

where $surface(A\_SF)$ is the surface of *A_SF* on which the relationship is docked. The value of ϕ is fixed by taking into account the needs of experts and in particular the size of the studied area.

• *Distance*: The indicators used in this relationship are of the form: "*x* km", "*x* miles", etc. Two

[3]National Institute of Geographic and Forestry Information - www.ign.fr

representations are then proposed in our approach: 1) calculus of distance from the centroid of the *A_SF* and construction of a circular buffer of size *x* from the centroid; 2) regarding the shape of the *A_SF* and building a buffer of size *x* from the edge of the processed *A_SF*.

• *Inclusion*: This binary operation allows us to check if an *A_SF* is inside another by taking into account indicators such as "center", "in the heart", "in", "inside", etc. Two types of inclusion are considered: 1) full inclusion, for instance, the expression "the town of Sete is in the Herault region"; 2) partial inclusion (i.e. intersection), for instance, "departmental road RD2 crosses (partly included) Poussan town".

• *Orientation*: This unary relationship has been broadly studied in the literature. Different approaches have been proposed to identify a cardinal points of an *A_SF*. We have chosen to use the conical model proposed in (Frank, 1991). For this, we use the centroid of *A_SF* and we build a buffer around. The size of this buffer will be calculated by taking into account the surface of the studied *A_SF*. Then we decompose the buffer into four equal areas (forming a *"X"*) from the centroid. Each intersection between the buffer and cones thus formed represent the four cardinal points.

• *Geometry*: Geometry relations are built from at least two *A_SF*. These relationships are, for example, the union, the adjacency, the difference or a position of an *A_SF* with respect to other *A_SF*, for example, C between A and B (where A,B and C are *A_SF*), etc.

Representation of the Footprint: After the extraction step and spatial representation of the *A_SF* and *R_SF*, the footprint associated with the treated document can be mapped. In this process, two main problems have been identified. The first one is the persistent ambiguity of some named entity contained in SF because some named entities correspond to several places, e.g. "Montagnac". To address this issue, a configurable spatial filter based on predefined scenarios has been developed. For example, to identify events related to a specific land-use planning project occurred in a part of the area of the Thau lagoon, only the SF contained in this area will be explored. The second issue is related to the use of external resources and the identification of the spatial representation appropriate to each *A_SF*. Taking into account the spatial indicator (e.g. "town", "road", etc.) preceding by the toponymic name is a first answer because it allows us to specify the type of the SF and thus take into account the appropriate spatial representation.

4 EXPERIMENTS

4.1 Data Preprocessing

A set of newspaper articles (i.e. 3809 documents) concerning the Thau basin from 2010 up to 2013 has been acquired. A second part of the data set is composed of three Pleiades satellite images (2 m pixel size and 4 spectral bands) covering the entire Thau basin region. Satellite images are available via the GEOSUD Equipex[4].

As cited before, in this article we propose a process to identify and construct footprints from textual documents to help the remote sensing analysts to explain changes. To achieve this, we need two inputs.

1) In one hand, we have to identify a spatial representation of a spatial object related to the land-use planning project that we want to address. This spatial representation is obtained thanks to a remote sensing process applied over a time series of satellite images. The goal is to find abrupt and unexplained changes in this time series. The three Pleiades images used in this work were acquired on 10 July 2012, 14 September 2012, and 15 March 2013 (see Figure 3).

Figure 3: Visible color composite of the three Pleiades images showing the construction of a bypass north of Villeveyrac (France).

First, a temporal clustering is performed over the Villeveyrac area (K-means algorithm with 12 classes). One of the obtained clusters (see Figure 4.a. - blue color) is related to the atypical change representing the bypass construction. Then, a segmentation (MeanShift algorithm) is applied over the temporal clustering results. In total, 1711 objects are obtained from the segmentation (see Figure 4.b.) and 8 of them correspond to the bypass construction. These 8 objects are exported as a single multi-polygon vectorial layer (see Figure 4.c.) and represent the footprint of the Villeveyrac bypass extracted from the Pleiades time series.

2) In the other hand, we have at our disposal an extract of our corpus containing 3809 newspapers articles provided by the French press Midi Libre. This corpus is related to the Thau basin region from 2010 up to 2013.

[4]http://www.equipex-geosud.fr/

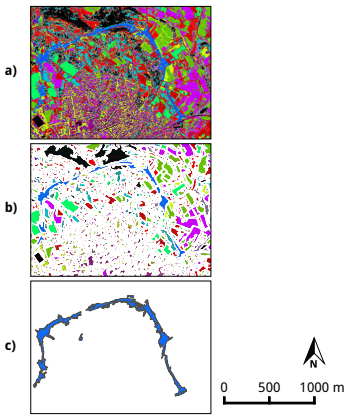

Figure 4: Clustering, segmentation and vector output from the Pleiades time series processing around the Villeveyrac area.

4.2 The Process to Match Images and Textual Documents

Once the textual corpus collected, the process, described in Section 3.2, has been applied to extract SF (see Table 1). There are two outputs files. The first one is A XML file, in which SF are identified and marked by a tag $<$ SF $>$. A SF is validated if it is present in the external geographical resources used (IGN and gazetteers).

In the XML file, one of the tags expresses a spatial relationship between one or two SF. Indeed, tag *indirection* represents one of the five categories of spatial relations described in Section 3. We can take advantage of *indirections* to build *R_SF*. To do this, all components in $<$ indirection $>$ tag are explored in order to extract information to extend *A_SF* to *R_SF*. In this manner, some SF associated to spatial relationships can be extended to approximate SF towards the spatial representation of the event identified in the use-case.

The second file (structured in JSON format) represents the geographical characteristics of confirmed SF. In fact, once SF is validated, geographical information describing the SF is collected from resources. This information is, for example, the Lambert coordinates, the resource name, times that SF appears in resources, etc. This information is presented following a JSON structure.

Figure 5 represents validated spatial entities extracted from 3809 documents. As presented in Table 1 (see line 1), 252575 spatial entities are identified and validated.

In this context, an important issue has to be tackled: a validated SF could represent different locations. For example, the SF "Croissiles" is a municipality located in 3 different areas in France: Pas-de-Calais, Calvados and Orne. In the same way, "Calais" is a town located in United States and as well in the north of France.

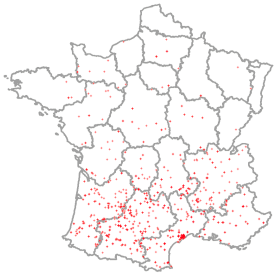

Figure 5: Spatial entities identified from corpus over the French territory.

To address this problem and reduce the noise associated to the ambiguity of spatial entities, we have filtered spatial entities located in the studied area (using Lambert coordinates). The treated experimental land-use planning project is the construction of the Villeveyrac bypass, which represents a small part of the Thau basin region. Thanks to this filtering step, we reduce by 95% potentially SF candidates (from 252575 points to 2090 SF, see Table 1, lines 1 and 2).

Nevertheless, our goal is to be as close as possible to an event that appeared in a specific area located thanks to our scenario. The representation using a simple point is not enough. In this work, we propose to take into account representation of *A_SF* and *R_SF* extracted by our process (see Section 3.2). For instance, the *A_SF "the town of Villeveyrac"* can be represented by a polygon (corresponding to territory of the municipality), the *A_SF "the road R2"* can be represented by lines (it is a departmental road), etc. These representations are obtained using external resources as IGN maps, GeoNames, etc. We have proposed a top-down geographic granularity level, from departments to buildings (in France). However, three types of representations have been considered for our experiments: municipalities, roads, and buildings because of the studied scenario that covers a relatively small area. Figure 6 shows 4 footprints, generated by our approach, which are in intersection relation with the spatial object representing the Villeveyrac bypass. As presented in Section 3.2, a footprint is obtained by the union of all validated SF extracted from a document.

We note that footprints are spatially closer to event previously identified in the representation of the treated spatial object (extracted from satellite images and related to the Villeveyrac bypass construction) thanks to Absolute and Relative Spatial Features (cf. *A_SF* and *R_SF*). Our method enables to identify a footprint for 754 textual documents, e.g. these documents contain at least one SF located in the studied area (see Table 1, line 4). From this selection, 279

Table 1: Experiments on a corpus of 3809 articles.

Steps	Nb Documents	Nb Points	Nb SF	Nb A_SF	Nb R_SF
1. Initial corpus	3809	252575	-	-	-
2. Filtering by the spatial area	758	2090	-	-	-
3. Intersection with points	0	0	-	-	-
4. Documents with footprints	758	-	2090	1992	98
5. Intersection with footprints	279	-	763	677	96

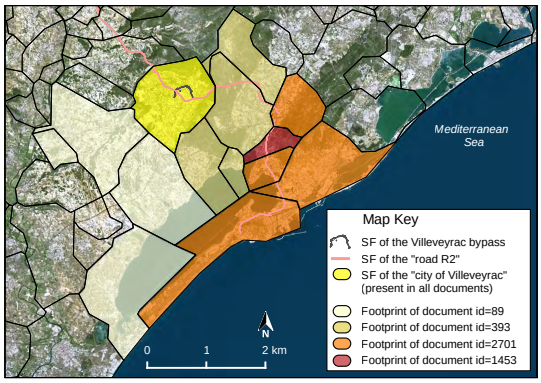

Figure 6: Representation of footprint related to 4 selected documents.

documents have got a footprint in intersection with the spatial object representing the Villeveyrac bypass (see Table 1, line 5).

In this work the footprints are identified, the next step is to analyze the document content to identify the ones referring to the event defined in the scenario.

5 CONCLUSION AND FUTURE WORK

This paper presents the first step of a methodology for extracting spatial footprint associated with textual documents. The main issue is to identify textual documents containing spatial footprint associated to satellite images, and this work is part of the ANIMITEX project. We have used the spatial relation of intersection to identify textual documents related to the same area of the spatial object representing the scenario. The second step will be performed via several spatial operators (i.e. distance, orientation, covering, etc.) according to a protocol which has to be fixed and in adequacy with the needs of the experts.

Finally, we propose to evaluate the reproducibility and relevance of experiments on other corpora and associated to a distinct land-use planning context.

ACKNOWLEDGEMENTS

The authors thank Midi Libre (French newspaper) for its expertise on the corpus, and all of the partners of ANIMITEX. This work is partially funded by Mastodons CNRS grant, and GEOSUD Equipex.

REFERENCES

Cohn, A. G. (1996). Calculi for qualitative spatial reasoning. *Artificial Intelligence and Symbolic Mathematical Computation*, 1138(5):124–143.

Davis, E. (2013). Qualitative spatial reasoning in interpreting text and narrative. *Spatial Cognition & Computation*, 13(4):264–294.

Forestier, G., Puissant, A., Wemmert, C., and Gançarski, P. (2012). Knowledge-based region labeling for remote sensing image interpretation. *Computers, Environment and Urban Systems*, 36(5):470–480.

Frank, A. (1991). Qualitative spatial reasoning with cardinal directions. *Seventh Austrian Conference on Artificial Intelligence*, 287:157–167.

Gaio, M. and Nguyen, V. (2011). Towards heterogeneous resources-based ambiguity reduction of sub-typed geographic named entities. In *Int. Conf. of GeoSpatial Semantics (GeoS)*, pages 217–234.

Hernández, D., Clementini, E., and Felice, P. (1995). Qualitative distances. *Spatial Information Theory A Theoretical Basis for GIS*, 988:45–57.

Hollenstein, L. and Purves, R. (2010). Exploring place through user-generated content: Using flickr tags to describe city cores. *J. Spatial Information Science*, 1(1):21–48.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11.

Lesbegueries, J., Gaio, M., and Loustau, P. (2006). Geographical information access for non-structured data. In *ACM Symposium on Applied Computing (SAC '06), ACM*, pages 83–89.

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol-Taravella, I., and Nouvel, D. (2011). Casen: a transducer cascade to recognize french named entities. *TAL*, 52(1):69–96.

Sallaberry, C., Gaio, M., and Lesbegueries, J. (2008). Fuzzying gis topological functions for gir needs. In *5th ACM Work. On Geog. Inf. Ret.*, pages 1–8.

Smith, D. A. and Mann, G. S. (2003). Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1*, pages 45–49.