Extraction et interprétation d'information géographique dans des données non-structurées

Julien Lesbegueries, Pierre Loustau

Catégorie Jeunes Chercheurs
Laboratoire LIUPPA - Thème IDEE
Université de Pau et des Pays de l'Adour
Avenue de l'Université
B.P. 1155
64013 Pau Cedex, FRANCE

julien.lesbegueries@univ-pau.fr
pierre.loustau@univ-pau.fr

RÉSUMÉ. *Cet article présente le projet "Pyrénées Itinéraires Virtuels". Ce projet consiste à valoriser un fonds documentaire patrimonial localisé dans le territoire pyrénéen. Dans ce cadre, nous proposons des modèles unifiés pour la définition formelle d'entités spatiales. Ces modèles permettent de mettre en place un système de recherche d'information basé sur le contenu sémantique de documents multi-formats. L'objectif de ce projet est d'étendre les fonctionnalités de systèmes de gestion de base documentaire classiques en permettant une gestion plus fine des restrictions spatiales dans une recherche. Pour cela nous développons un processus d'extraction d'information (EI) spécifique basé sur les modèles unifiés. De plus, une réflexion est menée sur l'interprétation numérique des entités spatiales. Un outil de recherche d'information (RI) utilise alors le traitement sémantique effectué pour retrouver des fragments de documents spatialement pertinents. Un prototype implémentant ce processus est développé afin de valider nos travaux.*

ABSTRACT. *This paper outlines the "Pyrénées Itinéraires Virtuels" project. The aim of this project is to add value to a legacy localised corpus. Unified models are proposed to define spatial entities in a formal way. These models allow to build a specific information retrieval system based on the semantic contents of various kinds of documents. Moreover, a reflection on computable interpretation of these spatial entities is performed, in order to be used in an information retrieval process. A prototype implementing this kind of information extraction and information retrieval process has been developped to validate our assumptions.*

MOTS-CLÉS : *raisonnement spatial, extraction et recherche d'information spatiale*

KEYWORDS: *qualitative spatial reasonning, information extraction, information retrieval*

1. Introduction

Les travaux présentés dans cet article sont effectués dans le cadre du projet "Pyrénées Itinéraires Virtuels"[1], dont le but est de valoriser un fonds documentaire patrimonial composé de documents hétérogènes : livres, journaux, cartes postales, lithographies, etc. Ces documents ne sont accessibles que dans les archives des musées ou des bibliothèques. Afin de permettre un accès plus généralisé, une campagne de numérisation et d'OCR-isation a été lancée. Cependant, ce processus n'est pas suffisant pour mettre en place un système performant de recherche d'information (RI), nécessaire à la resocialisation de ce corpus[CAS 04].

Ces documents ont la particularité d'être fortement attachés au patrimoine local, ceci nous a donc amené à concevoir un système de RI spécifique basé sur les entités géographiques. Pour cela nous avons élaboré un processus d'extraction d'information utilisant la sémantique des documents, plus fin qu'un processus utilisant les approches statistiques d'indexation. Afin d'indexer sémantiquement ces documents multi-formats, nous avons bâti un Modèle Spatial Unifié (MSU), permettant de fournir une représentation formelle pour ces données non-structurées.
Nous avons ensuite mis en place un moteur d'interprétation numérique de cette représentation afin d'utiliser les entités spatiales extraites dans un système de RI spatial.

2. Extraction de la sémantique spatiale des documents

L'information géographique contenue dans notre corpus est présente sous plusieurs modes d'expression. Chaque mode a ses spécificités. Si le texte est efficace pour décrire l'expression d'un phénomène sur un lieu géographique, une carte est plus appropriée lorsque l'on évoque l'organisation spatiale complexe d'un phénomène. Cependant, quel que soit le mode d'expression, cette information géographique apparaît sous forme d'Entités Géographiques (EGs). Ces dernières se composent d'une Entité Spatiale (ES), d'une Entité Temporelle (ET) qui peut être implicite et d'un phénomène. Afin de traiter correctement l'information géographique, une analyse fine des aspects spatiaux et temporels est obligatoire. Nous nous appuyons ici sur une approche sémantique qui a déjà fait ses preuves [CHA 03, WID 04].
Nous nous focaliserons dans cet article uniquement sur l'aspect spatial des EGs.

Le concept "cible/site" : Les travaux des linguistes montrent la manière particulière qu'a l'humain de se représenter une information spatiale lorsqu'elle est évoquée dans le langage écrit [BOR 98]. Faire référence à un lieu met en jeu plusieurs éléments et ces éléments respectent une position dans la phrase. [VAN 86] propose le concept de *cible/site* : dans le langage écrit, la cible correspond à l'objet de la description, le site à la référence. Notre hypothèse est d'étendre ce concept à d'autres modes d'expression tel que l'image par exemple.

1. Le projet PIV est soutenu par la Communauté d'Agglomération de Pau et la Médiathèque Intercommunale à Dimension Régionale de l'Aglomération Paloise.

Modèles unifiés : Afin de développer une Recherche d'Information efficace, nous proposons des modèles unifiés pour représenter le temps et l'espace. Contrairement à [EGE 02], [HIL 00] ou à GML[2] qui gère l'information spatiale du point de vue des bases de données, les modèles que nous proposons doivent supporter l'information géographique contenue dans notre corpus. Celle-ci est non structurée, polysémique, et dépendante du contexte.

A partir des travaux des linguistes, nous définissons une ES récursivement grâce à d'autres ES et à des relations spatiales (Figure 1). Prenons l'exemple de l'expression *"le nord de la ligne Pau-Biarritz"*. Dans cette expression l'ES évoquée est (i) tout d'abord définie par deux sites (ici deux entités nommées : *Biarritz* et *Pau*), (ii) puis le terme *ligne* exprime une relation géométrique linéaire entre ces deux sites. Enfin, (iii) *"le nord de"*exprime une orientation pour mettre en évidence l'espace évoqué.
Cette définition récursive peut aussi s'appliquer à d'autres modes d'expression. Dans une image, les entités nommées peuvent être représentées par des symboles ponctuels alors que des ES plus complexes peuvent s'évoquer par des relations entre ces ES nommées : une ligne puis deux zones contrastées par exemple.

Nous avons défini cinq relations spatiales qui sont l'adjacence, l'orientation, la distance, l'inclusion [COH 97] [COH 01] et la relation géométrique.

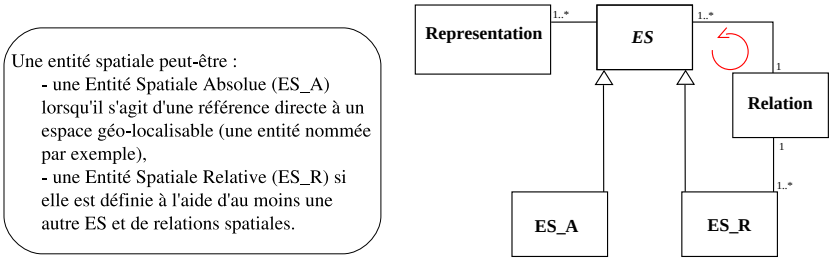

Figure 1. *Schéma UML simplifi é du Modèle Spatial Unifi é*

Analyse de la sémantique : Le processus d'Extraction de l'Information (EI) que nous préconisons ici se compose de quatre étapes principales : la lemmatisation, l'analyse lexicale et morphologique, l'analyse syntaxique et l'analyse sémantique [LES 06]. Ce processus d'EI a été validé pour les documents textuels par le développement de chaînes de traitement linguistique utilisant la plate-forme Linguastream[3] et des grammaires écrites en Prolog. Il produit des instances de MSU à chaque fois qu'une ES est détectée dans notre corpus.
Concernant les documents images, la validation des modèles a été faite manuellement mais il reste à automatiser la détection d'ES.

2. GML : Geography Markup Language (http ://opengis.net/gml)
3. Linguastream : http ://www.linguastream.org

3. D'une description qualitative des entités à leur représentation numérique

Le MSU défini dans la section précédente permet de garantir une homogénéité dans la description des entités spatiales, quel que soit le mode d'expression employé. En effet, nous parlons généralement de description et de raisonnement qualitatif de l'espace pour le mode d'expression textuel. Mais il existe aussi des travaux sur la description qualitative d'entités spatiales pour d'autres modes d'expression, comme dans une image par exemple [CAM 01]. Ce modèle unifié consiste donc en une première étape vers une représentation formelle de ces entités. Cependant, cette représentation générique n'est pas adaptée pour être utilisée directement dans un système de RI. Il est donc nécessaire de la traduire en une représentation plus propice à la comparaison et au classement, telle une représentation numérique et géo-référencée.
Le problème est alors de définir une méthode de traduction du MSU vers une telle représentation, dépendante du contexte et du degré de précision requis par le système de RI. Cette dernière représentation peut alors être utilisée comme comparateur et permet de calculer un degré de pertinence pour les entités lors du traitement d'une requête spatiale.

Utiliser une ontologie pour fabriquer des interprétations compréhensibles à la fois par l'homme et la machine est une idée assez répandue. [OCA 05] propose un modèle d'interprétation automatique de carte géographique, reposant sur un modèle de traduction basé sur une ontologie qui décrit le domaine d'étude (des objets géographiques pouvant être présents sur une carte). L'approche retenue dans notre cas est de pouvoir utiliser n'importe quelle ressource géographique (telle qu'une ontologie du spatial) pour fabriquer des interprétations spatiales pouvant être utilisées comme comparateurs dans un système de RI.

Problématique d'interprétations multiples : Il est important de considérer qu'une entité spatiale peut avoir de multiples représentations (figure 2). Par exemple une ville définie par une zone géométrique dans un SIG[4] peut être représentée par un point, un polygone, une boîte englobante, etc. La représentation est choisie pour un cas précis selon la base de connaissances employée ou selon les besoins. C'est pourquoi il est préférable de laisser la possibilité d'utiliser une base de connaissances moins riche ou simplifiée.
Par exemple, la masse de données spatiales stockée dans un SIG peut être très volumineuse et les calculs fins d'intersection ou de recouvrement des zones géométriques représentant les entités peuvent s'avérer coûteux. Nous pouvons alors utiliser des couches géographiques simplifiées. En outre, si l'on n'utilise pas de SIG pour la phase d'interprétation, nous devons pouvoir employer d'autres bases de connaissances tel un thésaurus décrivant un ensemble d'entités géographiques.

Une couche SIG pour ontologie : La communauté des SIG s'intéresse de plus en plus aux ontologies décrivant les phénomènes spatiaux [MAR 01]. L'association d'un SIG avec des ontologies décrivant les domaines du spatial est une pratique courante.

4. Système d'Information Géographique.

[FON 99] utilise les ontologies dans un système de gestion de données géographiques inter-opérables. Elles servent dans ce cas à assister la recherche d'information géographique et à faire inter-opérer plusieurs SIG ensemble.

Nous proposons ici de considérer un SIG et des couches de données géographiques comme base de connaissances servant à créer une représentation numérique des entités spatiales extraites de notre corpus. En effet, nous nous servons des concepts définis dans les couches de données d'un SIG, leur structure relationnelle et les fonctionnalités propres aux SIG pour interpréter nos entités décrites grâce au modèle unifié. Il en ressort pour chacune d'entre elles une zone géo-référencée manipulable par un moteur de recherche spatial. La figure 2 présente le modèle d'un tel système.
Dans le cas de l'utilisation d'un SIG, nous pouvons imaginer un format de sortie en GML pour l'interprétation numérique. D'autres formats peuvent aussi être envisagés, en XML, etc. Nous allons maintenant voir une implémentation de ce système développée pour valider ces principes.

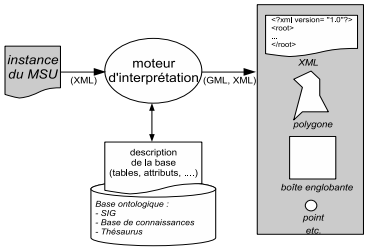

Figure 2. *Méthode d'interprétation multiple*

```
<entite_geographique id="4" id_paragraphe="2">
  <eg_r>
    <libelle>est d'Eaux-Bonnes</libelle>
      <relation>
        <orientation><type>est</type></orientation>
      </relation>
    <eg_a>
      <libelle>Eaux-Bonnes</libelle>
       <type>village</type>
    </eg_a>
  </eg_r>
  <representation>
    <boite_englobante>
      <x_min>360689.2</x_min><y_min>1752718.6</y_min>
      <x_max>389050.6</x_max><y_max>1789151.3</y_max>
    </boite_englobante>
  </representation>
</entite_geographique>
```

Figure 3. *Instance du MSU (en grisé l'ajout d'une représentation)*

Exemple d'utilisation d'un SIG : Dans le cadre du projet PIV a été développé un prototype de système d'extraction et de recherche d'information spatiale [LES 06]. Dans ce prototype, le moteur d'interprétation utilise un SIG contenant les communes de France comme base de connaissances. Prenons l'exemple d'une interprétation numérique du MSU. Dans un document se trouve l'entité *"à l'est d'Eaux-Bonnes"*. Celle-ci est extraite grâce à un traitement sémantique, puis une instance du MSU est créée. Ensuite le moteur d'interprétation, à l'aide d'une couche SIG, propose en sortie une boîte englobante représentant l'est d'Eaux-Bonnes (figure 3).
Cette représentation en boîte englobante est ensuite utilisée lors de la recherche d'information, permettant le calcul d'un degré de pertinence spatial.

4. Conclusion et travaux futurs

Nous présentons dans cet article les premiers travaux d'exploitation d'un corpus territorialisé à travers sa dimension spatiale. Un travail sur le raisonnement spatial a permis de concevoir une méthode de structuration de l'information géographique contenue dans les documents. Cette méthode a été validée par l'implémentation d'un prototype de RI spatiale.

Nous avons aussi mis en avant la nécessité de concevoir un système générique d'interprétation afin de pouvoir mettre en œuvre des algorithmes de comparaison spatiale. Cette interprétation doit pouvoir utiliser différentes ressources (ontologiques, géographiques, etc.) afin de s'adapter à différents processus de RI. La solution présentée dans cet article est une solution qui utilise un SIG, mais nous pourrions très bien définir d'autres interprétations utilisant des bases de connaissances plus classiques (ontologie, thésaurus, etc.).

En outre, un travail similaire sur l'aspect temporel doit être mené afin de compléter l'extraction d'information géographique des documents.

5. Bibliographie

[BOR 98] BORILLO A., *L'espace et son expression en français*, L'essentiel, Ophrys, 1998.

[CAM 01] CAMARA G., EGENHOFER M. J., FONSECA F. T., MONTEIRO A. M. V., « What's in an Image ? », *Spatial Information Theory*, 2001, p. 474-488.

[CAS 04] CASENAVE J., MARQUESUZAÀ C., DAGORRET P., GAIO M., « La revitalisation numérique du patrimoine littéraire territorialisé », *EBSI-ENSSIB, Montréal*, 2004.

[CHA 03] CHARNOIS T., MATHET Y., ENJALBERT P., BILHAUT F., « Geographic Reference Analysis for Geographic Document Querying », Workshop of the NAACL-HLT Conference, Association for Computational Linguistic, 2003.

[COH 97] COHN A. G., « Qualitative Spatial Representation and Reasoning Techniques », *KI '97 : Proceedings of the 21st Annual German Conference on Artificial Intelligence*, London, UK, 1997, Springer-Verlag, p. 1–30.

[COH 01] COHN A. G., HAZARIKA S. M., « Qualitative Spatial Representation and Reasoning : An Overview », *Fundamenta Informaticae*, vol. 46, n° 1-2, 2001, p. 1–29.

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

[FON 99] FONSECA F. T., EGENHOFER M. J., « Ontology-driven geographic information systems », *GIS '99 : Proceedings of the 7th ACM international symposium on Advances in geographic information systems*, New York, NY, USA, 1999, ACM Press, p. 14–19.

[HIL 00] HILL L. L., « Core Elements of Digital Gazetteers : Placenames, Categories, and Footprints », *ECDL '00 : Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, Springer-Verlag, 2000, p. 280–290.

[LES 06] LESBEGUERIES J., GAIO M., LOUSTAU P., SALLABERRY C., « Geographical information access for non-structured data », ACM ASIIS - Dijon - à paraître, 2006.

[MAR 01] MARK D., EGENHOFER M., HIRTLE S., SMITH B., « Ontological Foundations for Geographic Information Science », 2001.

[OCA 05] MONTES DE OCA MORALES V., « Template-Based Geospatial Knowledge Representation », Short Paper Geos2005, 2005.

[VAN 86] VANDELOISE C., *L'espace en français*, 1986.

[WID 04] WIDLÖCHER A., FAUROT E., BILHAUT F., « Multimodal Indexation of Contrastive Structures in Geographical Documents », *Actes RIAO 2004, Avignon*, 2004, p. p. 555-570.