Actes de l'atelier Recherche d'Information SEmantique, RISE 2015

J.P. Chevallet, Catherine Roussey, Haifa Zargayouna

Septième Atelier Recherche d'Information SEmantique RISE, Rennes 30 juin 2015

Associé à IC@PFIA 2015

Actes de l'Atelier Recherche d'Information SEmantique RISE 2015

Édité par
Jean-Pierre CHEVALLET, LIG, Grenoble (France)
Catherine ROUSSEY, IRSTEA, Clermont Ferrand (France)
Haïfa ZARGAYOUNA, LIPN, Paris (France)

Septième édition Atelier Recherche d'Information SEmantique

Atelier Recherche d'Information SEmantique RISE, Rennes 30 juin 2015

Associé à IC@PFIA 2015

1. Introduction

Nous avons le plaisir d'organiser à Rennes la septième édition de l'atelier Recherche d'Information SEmantique, RISE 2015, associé à la conference Ingénierie des Connaissances de la Plateforme d'Intelligence Artificielle et avec le soutien de l'ARIA (Association francophone de Recherche d'Information et Applications).

Le but de l'atelier est de proposer un espace d'échange autour de la synergie entre acquisition et gestion de ressources sémantiques (ontologies, terminologies, thesaurii, ...) et la Recherche d'Information. Ces thématiques sont à la croisée du Web Sémantique, de l'Ingénierie des Connaissances, du Traitement Automatique des Langues et de la Recherche d'Information.

Nous avons le plaisir cette année d'accueillir Clément Jonquet pour la conférence invitée qui a pour titre «A few contributions of the SIFR (Semantic Indexing of French biomedical Resources) project and how we reuse NCBO technology.».

Clement Jonquet est maître de conférences en Informatique à l'Université de Montpellier. Il est chercheur au Laboratoire d'Informatique, Robotique et de Microélectronique de Montpellier (LIRMM), sur les domaines des ontologies (biomédicales), indexation sémantique de données, l'annotation, le Web sémantique, la fouille de texte et les questions de représentation des connaissances. Clement Jonquet est docteur en informatique de l'Université Montpellier 2 depuis 2006 et a réalisé un post-doctorat de 3 ans à l'Université de Stanford, dans l'équipe de Mark A. Musen, où dans le cadre du projet National Center for Biomedical Ontology (NCBO) il a participé activement à la mise en œuvre du NCBO BioPortal, le portail d'ontologies de référence en biomédecine. Depuis 2013, il est le coordinateur du projet SIFR (Semantic Indexing of French Biomedical Data Resources) financé principalement par le programme JCJC de l'ANR. C. Jonquet est (co)auteur de +42 publications, qui cumulent plus 1300 citations. Depuis 2010, il est enseignant à Polytech' Montpellier dans le département Informatique et Gestion. Il est responsable des technologies de l'information et de la communication pour l'enseignement (TICE) pour l'école d'ingénieur.

Trois doctorants sont également invités à présenter leur sujet de thèse lors d'une session dédiée. Nous remercions vivement les auteurs, les membres du comité de programme ainsi que les participants qui assurent chaque année la bonne tenue de l'atelier.

2. Comité de programme

- BELLOT Patrice, LSIS Avignon (France)
- BERTIN Marc, STIH Paris (France), CIRST Montreal (Canada)
- BUSCALDI Davide, LIPN, Paris (France)
- CALABRETTO Sylvie, LIRIS Lyon (France)
- CHEVALLET Jean-Pierre, LIG, Grenoble (France)
- GRAU Brigitte, ENSIIE (France)
- HERNANDEZ Nathalie, IRIT Toulouse (France)
- KAMEL Mouna, IRIT Toulouse (France)
- ROUSSEY Catherine, IRSTEA, Clermont Ferrand (France)
- SALOTTI Sylvie, LIPN, Paris (France)
- SCHWAB Didier , LIG-GETALP, Grenoble (France)
- SERASSET Gilles, LIG, Grenoble (France)
- TAMINE LECHANI Lynda, IRIT, Toulouse (France)
- ZARGAYOUNA Haïfa , LIPN, Paris (France)

3. **Table des matières**

A few contributions of the SIFR (Semantic Indexing of French biomedical Resources) project and how we reuse NCBO technology

Clement JONQUET

LIRMM, Université de Montpellier & CNRS
jonquet@lirmm.fr

Résumé : In this talk I will quickly present the research and development contributions of the SIFR project which investigates the scientific and technical challenges of developing ontology-based services to leverage the use of biomedical ontologies and terminologies for indexing, mining and retrieval of French biomedical data. Within this project we extensively reuse and complement NCBO technology with the goals of:

- Designing and implementing a French version of the NCBO Annotator web service (the French Annotator) as well as implementing new features for the annotation workflows.
- Designing and implementing a multilingual version of BioPortal which can handle different kind of multilingual ontologies and alignments.
- Deploying a specific instance of the portal for the agronomic, plant and environment domain (the AgroPortal) in collaboration with IRD, CIRAD, INRA and Bioversity.

I will present our local developments and walk through the main research contributions of the project in terms of automatic term extraction, semantic annotation, multilingual representation, semantic distances, informal patient data analysis, semantic indexing of data. I will present how each of these contributions are somehow relevant to the topics of RISE workshop and semantic search.

Mots-clés : Modèles de Recherche d'Information Sémantique, Extraction d'Information, Annotation Sémantique, Indexation Sémantique, Alignement d'ontologies et correspondances pour la Recherche d'Information, Langages de Représentation des connaissances pour la Recherche d'Information, Utilisation des distances Sémantiques pour la Recherche d'Information.

Références :

Soumia Melzi & Clement Jonquet. **Scoring semantic annotations returned by the NCBO Annotator**, In *7th International Semantic Web Applications and Tools for Life Sciences, SWAT4LS'14*. Berlin, Germany, December 2014. CEUR Workshop Proceedings, Vol. 1320 pp. 15. CEUR-WS.org.

Clement Jonquet, Vincent Emonet & Mark A. Musen. **Roadmap for a multilingual BioPortal**, *In 4th Workshop on the Multilingual Semantic Web, MSW4*. Portoroz, Slovenia, June 2015

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche & Maguelonne Teisseire. **Towards a mixed approach to extract biomedical terms from text corpus**, *Knowledge Discovery in Bioinformatics*. 2014. Vol. 4 (1), pp. 15. IGI Global.

Guillaume Surroca, Philippe Lemoisson, Clement Jonquet & Stefano A. Cerri. **Diffusion de systèmes de préférences par confrontation de points de vue, vers une simulation de la Sérendipité**, In *26èmes Journées Francophones d'Ingénierie des Connaissances, IC'15*. Rennes, France, June 2015.

Guillaume Surroca, Philippe Lemoisson, Clement Jonquet & Stefano A. Cerri. **Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d'information scientifique**, In *25èmes Journées Francophones d'Ingénierie des Connaissances, IC'14*. Clermont-Ferrand, Mai 2014. pp. 12.

Clement Jonquet, Adrien Coulet, Nigam H. Shah & Mark A. Musen. **Indexation et intégration de ressources textuelles à l'aide d'ontologies : application au domaine biomédical**, In *21èmes Journées Francophones d'Ingénierie des Connaissances, IC'10*. Nimes, France, June 2010. pp. 271-282.

Crédits & Remerciements :

Ce travail est supporté par le projet by *Semantic Indexing of French biomedical Resources* (SIFR – www.lirmm.fr/sifr) financé en partie par le programme JCJC de l'Agence nationale de la Recherche (ANR-12-JS02-01001), l'Université de Montpellier, le CNRS, l'Institut de Biologie Computationnelle de Montpellier (ANR-11-BINF-0002) et le Labex Numev (ANR-10-LABX-20). Nous remercions le US National Center for Biomedical Ontology (NCBO) pour leur disponibilité et la mise à disposition de BioPortal.

Ce travail est le résultat d'une contribution de diverses personnes à Montpellier : Mathieu Roche, Sandra Bringay, Stefano A. Cerri, Maguelonne Teisseire, Pascal Poncelet, Vincent Emonet, Juan-Antonio Lossio-Ventura, Guillaume Surroca, Philippe Lemoisson, Pierre Larmande.

Machine Reading for Abstractive Summarization of Customer Reviews in the Touristic Domain

Ehab Hassan[1], Davide Buscaldi[1], Aldo Gangemi[1,2]

[1] LIPN CNRS UMR 7030, Laboratoire d'Informatique de Paris Nord, Villetaneuse, France
{ehab.hassan,davide.buscaldi}@lipn.univ-paris13.fr

[2] STLab, ISTC-CNR, Rome, Italy
aldo.gangemi@istc.cnr.it

Abstract : Abstractive summarization is the task of producing a concise representation from a more complex text or a set of texts. This is a useful task especially in the summarization of customer reviews. In this paper we present an abstractive summarization method based on a machine reader and sentiment analysis dictionaries. We carried out a preliminary evaluation of the method on 15 hotel reviews from the Opinosis collection.

Mots-clés : Machine Reading, Abstractive Summarization, Opinion Analysis

1 Introduction

Text summarization is a task consisting in the production of a concise description of a longer, more complex text. Usually, summarization approaches can classified into two types: *extractive* and *abstractive*. In the first case, the original text is reduced to a smaller one, keeping the most important fragments. In the latter, a new text is produced on the basis of the context of the original one. Therefore, abstractive summarization needs a deeper comprehension of the underlying semantics, where extractive summarization can be considered as a shallower task, where the semantics does not play an important role.

One of the most recent applications of the abstractive approaches is the summarization of product reviews and opinions Ganesan *et al.* (2010). This is particularly useful in cases where there are many reviews and most of them are redundant: a user may have to read a great quantity of text before being able to obtain a precise idea of the qualities and the disadvantages of a product.

Machine readers have been introduced by Etzioni *et al.* (2006) as tools for text understanding. They combine different text analysis layers (Part-Of-Speech tagging, syntactic analysis, disambiguation, named entity recognition) to produce a rich semantic representation of the text, which is the reason why we chose to apply them to user reviews in the touristic domain for the abstractive summarization of opinions.

The rest of the paper is structured as follows: in Section 2 we describe the process to extract the opinions and the features from the user reviews. In Section 3 we describe the summarization steps, while in Section 4 we show the experiments carried out and the obtained results. Finally, in Section 5 we draw some conclusions about this preliminary work.

2 Features and Opinion Extraction

The first step consists in the identification of features (or aspects) that are the object of evaluation by users. When users write a review of an hotel, for instance, they usually evaluate not the hotel in its entirety, but specific features of the hotel. Then, we need to find the attributes used to express the opinions. We assumed that such attributes are usually expressed as adjectives. In order to extract the features with their associated attributes from the user reviews, we perform a deep semantic parsing of text, obtaining a RDF Linked-Data-ready graph representation of the text. We employ a large variety of machine reading systems, as implemented in the FRED tool [1] Presutti *et al.* (2012), which extracts knowledge (named entities, senses, taxonomies, relations, events) from text, resolves it onto the Web of Data, adds data from background knowledge, and represents all that in RDF and OWL.

FRED is a tool to automatically transform knowledge extracted from text into RDF and OWL, i.e. it is a *machine reader* for the Semantic Web. It is available as a RESTful API and as a web application. In its current form, it relies upon several NLP components: Boxer[2] for the extraction of the basic logical form of text, BabelNet Navigli & Ponzetto (2010) for word sense disambiguation, and Apache Stanbol[3] for named entity resolution.

Since review features are usually nouns or noun phrases in user reviews, we only interested in features that appear explicitly as nouns or nouns phrases in the reviews. Applying a SPARQL query to the semantic graph produced by FRED, we can extract these features with their opinion words:

```
PREFIX dul: <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#>
PREFIX vnrole: <http://www.ontologydesignpatterns.org/ont/vn/abox/role/>
PREFIX boxing: <http://www.ontologydesignpatterns.org/ont/boxer/boxing.owl#>
PREFIX boxer: <http://www.ontologydesignpatterns.org/ont/boxer/boxer.owl#>
PREFIX : <http://www.ontologydesignpatterns.org/ont/boxer/test.owl#>
PREFIX d0: <http://www.ontologydesignpatterns.org/ont/d0.owl#>
PREFIX schemaorg: <http://schema.org/>
PREFIX fred: <http://www.ontologydesignpatterns.org/ont/fred/domain.owl#>
PREFIX pos: <http://www.ontologydesignpatterns.org/ont/fred/pos.owl#>
PREFIX BE: <http://www.essepuntato.it/2008/12/earmark#>
SELECT  distinct ?Feature ?neg ?qlt
WHERE {
{
        {?Feature rdf:type ?FeatureType}.
        {{?FeatureType pos:boxerpos pos:n}
          UNION{?FeatureType rdfs:subClassOf* ?FeatureType_1.?FeatureType_1 pos:boxerpos pos:n}}.
          {{?FeatureType dul:hasQuality ?qlt}
            UNION{?Feature dul:hasQuality ?qlt}}
}
OPTIONAL {?sit boxing:involves ?Feature . ?sit boxing:involves ?qlt . ?sit boxing:hasTruthValue ?neg}
}
```

This SPARQL query allows to extract noun features with their modifiers e.g. logical negations, and adverbial qualities (opinion words). Logical negations are very important to determine the polarity of features.

3 Summarization

The system performs the summarization in three main steps: (1) - identify features that have been commented on by customers; (2) - identify opinion words and their polarity, and deciding whether each opinion word is positive, negative, or neutral; and (3) - summarize the results

[1] http://wit.istc.cnr.it/stlab-tools/fred

[2] http://svn.ask.it.usyd.edu.au/trac/candc/wiki/boxer

[3] http://stanbol.apache.org

using the redundant opinions. Given an input composed by a set of user reviews, the system first extracts all the features that appear explicitly as nouns or noun phrases in the reviews and have at least one opinion word associated with them, together with their attributes and eventually the associated logical negation. Then, we used three sentiment lexicons (SentiWordNet (Baccianella *et al.* (2010)), AFINN[4] and Liu (2012)) to detect the polarity of the opinion words. We had three types of polarity which can be assigned to opinion word (e.g. positive, negative, or neutral). Afterward, we kept the features that have positive and negative polarity and deleted the neutral ones. To generate the final summary, we regrouped the remaining features by measuring the similarity between them. We used the WordNet:Similarity package by Pedersen *et al.* (2004) to measure the similarity between the features and considered that two features can be grouped together (i.e. consider as synonyms) if their Lin similarity score is greater than 0.5. We didn't take into account this method to group together the attributes since the WordNet:Similarity package does not offer good semantic similarity measures for adjectives (the only one is the Lesk measure which is not as reliable as the Lin one).

4 Experiments and Results

We started with 15 reviews of a randomly picked hotel from the Opinosis collection by Ganesan *et al.* (2010). We analyzed the reviews with FRED, extracting 140 attributed features. In Figure 1 we show a subset of the features retrieved from the 15 reviews and the associated attribute/opinion word.

Area	Clean	Hotel	Warm	Room	Great
Area	Nice	Location	Not_Better	Room	Huge
Area	Pleasant	Location	Excellent	Room	Large
Bathroom	Nice	Location	Excellent	Room	Spacious
Bathroom	Spacious	Location	Great	Room	Spotless
Bathroom	Spacious	Location	Perfect	Roomy	Clean
Bed	Comfortable	Location	Perfect	Service	Excellent
Bed	Perfect	Location	Right	Service	Great
Bed	Quiet	Person	Happy	Staff	Friendly
Door	Fantastic	Person	Nervous	Staff	Helpful
Hallway	Dark	Person	Picky	Staff	Helpful
Hallway	Loud	Person	Pleased	Stay	Enjoyable
Hotel	Not_Excellent	Rate	Awesome		
Hotel	Big	Rate	Excellent		
Hotel	Friendly	Rate	Great		
Hotel	Great	Rate	Great		
Hotel	Great	Restaurant	Incomplete		
Hotel	Great	Restaurant	Unavailable		
Hotel	Nice	Room	Comfortable		
Hotel	Upscale	Room	Excellent		

Figure 1: An excerpt of the 140 attributed features extracted from the 15 reviews. The highlighted features have been grouped together on the basis of their WordNet::Similarity distance.

The next step was to find the polarity of each attribute. For instance, "comfortable" has a positive polarity in all three dictionaries, and "nervous" has negative polarity in all dictionar-

[4]https://github.com/abromberg/sentiment_analysis/tree/master/AFINN

ies. We reduced the three polarities to a single value and then found the redundant attributed features. Therefore, the 15 reviews were summarized to the attributed features in Table 1.

Feature	Attribute	Freq	Polarity	Feature	Attribute	Freq	Polarity
Staff	Helpful	2	+	Hotel	Great	3	+
Location	Perfect	2	+	Location	Excellent	2	+
Rate	Great	2	+	Rate	Bad	2	-
Room	Spacious	3	+	Room	Nice	2	+
Bed	Comfortable	2	+				

Table 1: The result of the summarization of the 15 reviews.

5 Conclusions

Although this is a very preliminary work, we were able to reduce effectively the complete set of opinion to a synthetic table of features and attributes. Further directions may be to combine the attributes that are very similar (“perfect", “excellent"), using semantic similarity measures developed for Semeval STS[5], and find a way to deal with conflicting ratings. We need also to carry out a more comprehensive evaluation and compare to other summarization methods, such as the one proposed by Popescu & Etzioni (2007).

References

BACCIANELLA S., ESULI A. & SEBASTIANI F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, p. 2200–2204.

ETZIONI O., BANKO M. & CAFARELLA M. (2006). Machine reading. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*.

GANESAN K., ZHAI C. & HAN J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, p. 340–348: Association for Computational Linguistics.

LIU B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, **5**(1), 1–167.

NAVIGLI R. & PONZETTO S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, p. 216–225: Association for Computational Linguistics.

PEDERSEN T., PATWARDHAN S. & MICHELIZZI J. (2004). Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, p. 38–41: Association for Computational Linguistics.

POPESCU A.-M. & ETZIONI O. (2007). Extracting product features and opinions from reviews. In A. KAO & S. POTEET, Eds., *Natural Language Processing and Text Mining*, p. 9–28. Springer London.

PRESUTTI V., DRAICCHIO F. & GANGEMI A. (2012). Knowledge extraction based on discourse representation theory and linguistic frames. In *EKAW: Knowledge Engineering and Knowledge Management that matters*.

[5]http://alt.qcri.org/semeval2015/task2/

Entity-Based Information Retrieval System: A Graph-Based Document and Query Model

Mohannad ALMASRI[1], Jean-Pierre Chevallet[2], Catherine Berrut[1]

[1] UNIVSERSITÉ JOSEPH FOURIER - GRENOBLE 1, LIG laboratory, MRIM group, Grenoble, France
mohannad.almasri@imag.fr catherine.berrut@imag.fr

[2] UNIVERSITÉ PIERRE-MENDÈS-FRANCE - GRENOBLE 2, LIG laboratory, MRIM group, Grenoble, France
jean-pierre.chevallet@imag.fr

Abstract : Named Entity has been playing an important role in information seeking and retrieval. Many queries in web search are related to entities. Several knowledge bases contain valuable information about named entities and their relations, such as Wikipedia, Freebase, DBpedia and YAGO. Most existing works, about entity-based search, propose exploiting knowledge about entities and their relationships for expanding or reformulating a user query. Query expansion and reformulation yield effective retrieval performance on average, but results a performance inferior to that of using the original query for many information needs. In this paper, we propose to differently investigate knowledge about named entities and their relations. Therefore, we first present an entity-based document and query model. Then, we suggest a retrieval model based on language models to match between document and query models.

Keywords: Knowledge Base, Named Entity, Information Retrieval, Language Models.

1 Introduction

Named Entity plays an important role in information seeking and retrieval. According to some statistics, about 71% of queries in web search contain named entities (Guo *et al.*, 2009). knowledge bases like Wikipedia, Freebase, DBpedia and YAGO, contain valuable information about entities and their relations. These knowledge bases are essentially used for identifying entities in a user query. In an entity-based search scenario, retrieval systems exploit relationships between entities in order to expand or reformulate a named entity query. (Liu *et al.*, 2014; Dalton *et al.*, 2014; Audeh *et al.*, 2014; ALMasri *et al.*, 2013; Guo *et al.*, 2009; Xu *et al.*, 2008).

Query expansion yields effective retrieval performance on average, but results a performance inferior to that of using the original query for many information needs[1] (Zighelnic & Kurland, 2008). We think that one of the reasons that yields to this problem in entity-based search that the related entities are integrated as keywords into the original query. For example, the query «Silent Film» which searches for documents on history of silent film, actors and directors. A document talks about «Charlie Chaplin», for instance, is a relevant document to this query. However, in a classical query expansion system, the term «Charlie Chaplin» is added to the original query «Silent Film» as two keywords: «Charlie» and «Chaplin». As a result, a classical retrieval model retrieves documents contain «Charlie», «Chaplin», and «Charlie Chaplin» without respecting that «Charlie Chaplin» represents in total one entity.

[1] This robustness issue is called the query drift problem.

Our main contributions, in this study, are three fold:

- Identifying named entities in documents as well as in queries using a knowledge-based for entity detection.
- Proposing an entity-based a document and a query model starting from the identified entities and their relationships.
- Proposing a retrieval model based on language models framework (Ponte & Croft, 1998) for matching between our document and query models, and takes into account entity relations.

An entity is something that has a distinct, separate existence, which can be a person, a place, an organization or miscellaneous. The information associated with an entity is more abundance and less ambiguous for retrieval task than query keywords. Thus, it is our intuition that retrieval performance can benefit from passing into an entity-based retrieval system.

We use a knowledge-based approach for entity detection in documents and queries, where we propose to use Wikipedia as a knowledge repository about named entities. Wikipedia is a free online encyclopedia, it records one article for a real world entity, with information focuses on this entity. Wikipedia contains a huge number of linked articles about named entities. It is in fact a large manually edited repository of entities. Its large volume of structured data, and high quality content make it a convenient and perfect knowledge source which could play an important role in named entity detection.

2 Entity-Based Search System

Three essential components are existed in an information retrieval system: a document model, a query model, and a retrieval model that matches document and query model. Given a query and a document, we identify entities which are mentioned within them based on Wikipedia. Then, we build an entity-based graph representation for a query and a document. Finally, We adapt language models to achieve the matching between document and query graphs. We detail these steps in the following.

2.1 Wikipedia as a Graph

Wikipedia is an encyclopedia that represents a very large, high quality, and valuable knowledge source in natural language. Moreover, Wikipedia is also a hypertext in which each Wikipedia article can refer to other Wikipedia articles using hyperlinks. We consider only *internal links*, i.e. links that target another Wikipedia article.

We represent Wikipedia as a directed graph $G(A, L)$ of articles A, connected by links $L \subseteq A \times A$. Each article $a \in A$ is a description of an object, an entity, an historical fact, etc.. Furthermore, each article contains links to other articles. Relations between articles L are defined on $A \times A$, where (a_1, a_2) means that the article a_1 shows a link to the article a_2. In this article, we define:

$$\begin{array}{rcl} I, O & : & A \rightarrow 2^A \\ I(a) & = & \{x \in A | (x, a) \in L\} \\ O(a) & = & \{x \in A | (a, x) \in L\} \end{array}$$

where $I(a)$ is the set of articles that point to a (*Incoming Links*) , and $O(a)$ is the set of articles that a points to (*Outgoing Links*).

We propose to weight the Wikipedia graph in order to evaluate the strength of links between articles. For that, we consider that two Wikipedia articles are semantically similar, if they share similar links, i.e. if they have similar incoming and outgoing link sets.

Hence, two articles a_1 and a_2 in A are semantically similar if they share articles that point to them, and if they share articles that a_1 and a_2 point to. Then, we propose the following semantic similarity:

$$SIM(a1, a2) = \frac{|I(a_1) \cap I(a_2)| + |O(a_1) \cap O(a_2)|}{|I(a_1) \cup O(a_1)| + |I(a_2) \cup O(a_2)|} \quad (1)$$

where $I(a)$ are incoming links to article a, and $O(a)$ are outgoing links from a.

2.2 Document and Query Models

- **Entity Detection**. Given an n-word query q (w_1, w_2, ..., w_n). Instead of representing the query q by their keywords, we represent the query by entities which are mentioned within. Query annotation establish a link from query sub-phrases to entities in a knowledge base. To do this, we verify for each sub-phrase of q if there is an entity page entitled exactly by this sub-phrase, or it is redirecting to an entity page. Figure 1 shows a sentence that contains four named entities: silent film, film, recorded sound, and dialogue. Similarly, given a document d, we apply the same annotation strategy for each sentence in this document. Finally, we obtain a list of entities for each document or query.

A silent film is a film with no synchronized recorded sound, especially with no spoken dialogue.

Figure 1: Example of a sentence that contains four named entities: silent film, film, recorded sound, and dialogue.

- **Entity Linking**. Following the detection step, where each document or query is represented as a list of entities which are mentioned within. In Wikipedia, each entity page is linked with hyperlinks to a number of other entity pages. We inherit these links into our document and query representation, i.e. we link between two entities in a document or a query representation if there is a link between these two entities in Wikipedia. As a result, documents and queries are represented as a sub-graph of Wikipedia graph. Figure 2 gives an example of document and query representation.

2.3 Retrieval Model

In the previous section, we see that each document and query are represented by a graph of named entities, we look for a retrieval model achieving the following two goals:

– **Non-Matching Entities**. First goal is to deal with unmatched query entity, i.e. query entity which does not appear in a document, e.g. like e_2 in figure 2. In this case, we verify the existence of any link between this entity and other document entities in order to reduce the semantic gap during the matching between a document model d and a query model q (considering dashed links in figure 2).
– **Entity-Centered Documents**. Second goal is to identify from a series of documents that mention a given query entity, those which are entity-centered. Intuitively, we assume that an entity-centered document contains entities linked to this entity. For instance, the case of the query «Silent Film», a document contains the two linked entities «Silent Film» and «Charlie Chaplin», should be ranked before another document contains «Silent Film» with another non-linked entity like «Lionel Richie».

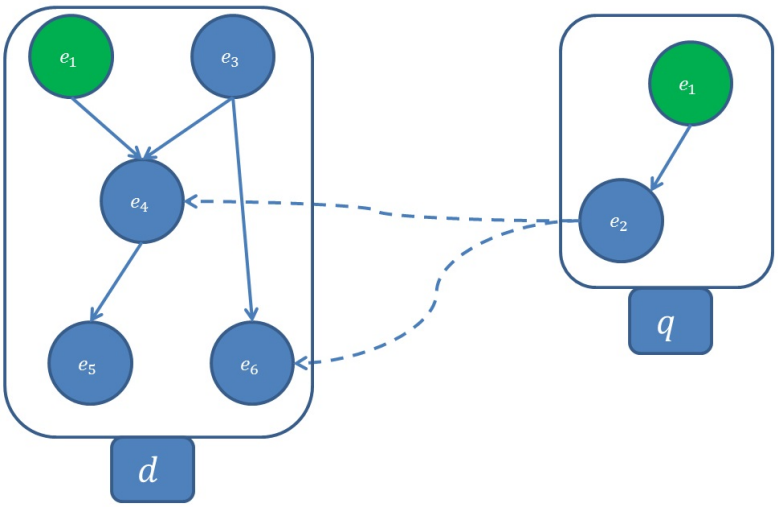


Figure 2: Example of document and query model. The document d contains five entities (e_1,e_3,e_4,e_5,e_6). The query q contains two linked entities (e_1,e_2).

In our approach, we propose to model these two concepts using the language models framework (Ponte & Croft, 1998).

2.3.1 Language Models in Information Retrieval

The basic idea of language models assumes that a query q is generated by a probabilistic model based on a document d. Language models are interested in estimating $P(d|q)$, i.e. the probability that a document d is used to generate a query q. By applying Bayes' formula, we have:

$$P(d|q) \propto P(q|d).P(d) \tag{2}$$

$\propto$ means that the two sides give the same ranking. $P(q|d)$ the query likelihood for a given document d. $P(d)$ is often assumed to be uniform, and thus it is discarded for ranking docu-

ments. We can rewrite $P(q|d)$ the query likelihood after adding the log function as:

$$logP(q|d) = \sum_{e \in V} \#(e;q).logP(e|d) \tag{3}$$

where $\#(e;q)$ is the count of entity e in the query q and V is the vocabulary set. Assuming a multinomial distribution, the simplest way to estimate $P(e|d)$ is the maximum likelihood estimator:

$$P_{ml}(e|d) = \frac{\#(e;d)}{|d|} \tag{4}$$

where $|d|$ is the document length. Due to the data spareness problem, the maximum likelihood estimator directly assign $null$ to the unseen entities in a document. Smoothing is a technique to assign extra probability mass to the unseen entities in order to solve this problem.

Jelinek-Mercer smoothing Zhai & Lafferty (2004) is one of the smoothing technique to add an extra pseudo entity frequency $\lambda P(e|C)$, based on the collection entity collection, to the unseen entity as follows:

$$P_\lambda(e|d) = (1-\lambda)P(e|d) + \lambda P(e|C) \tag{5}$$

We distinguish in the previous equation two main parts: a part related to the document $P(e|d)$, and another part related to the collection $P(e|C)$. In fact, the next section provides our adaption of language models which is based on modifying the way we estimate the document part probability $P(e|C)$ in order to achieve our two goals.

2.3.2 Language Models Adaptation

Our approach to achieve our two goals: **Non-Matching Entities** and **Entity-Centered Documents**, is based on modifying the way we estimate the probability $P(e|d)$ inside the the language models framework.

- **Non-Matching Entities**. We aim to reduce the semantic gap during the matching between a document model d and a query model q. To do this, we propose to modify a document model according to the query and the external knowledge about entity relations.

  Classical IR models compute the relevance value between a document d and a query q based on the coordination level, namely $d \cap q$. Instead of that, we here propose to compute the relevance value by considering also the *unmatched entities* of the query $e \in q \setminus d$, where $\setminus$ is the set difference operator. We therefore expand d by the query entities that are not in the document, but they are semantically linked to at least one document entity (like e_2 in figure 2). In this way, we maximize the coordination level between the document and the query. As a result, we maximize the probability of retrieving relevant documents for a given query. To put it more formally, the modified document, denoted by d_q, is calculated as follows:

  $$d_q = d \cup F(q \setminus d, G, d) \tag{6}$$

  where $F(q \setminus d, G, d)$ is the transformation of $q \setminus d$ according to the knowledge graph G and the document d. The knowledge graph G provides a similarity function between entities

$SIM(e, e')$, see formula 1, denoting the strength of the semantic relatedness between the two entities e and e'. For each entity $e \in q \setminus d$, we look for a document entity e^* which is given by:

$$e^* = argmax_{e' \in d} SIM(e, e') \tag{7}$$

e^* is the most similar entity of d for $e \in q \setminus d$. Then, the pseudo frequency of a query entity e in the modified document d_q relies on the frequency of its most similar document entity $\#(e^*; d)$, we define the pseudo frequency of e as follows:

$$\#(e; d_q) = \#(e^*; d).SIM(e, e^*) \tag{8}$$

This pseudo frequency of the entity e is then included into the modified document d_q. Based on this definition, we now define the transformation function F which expands the document.

$$F(q \setminus d, G, d) = \{e | e \in q \setminus d, \exists e^* \in d, e^* = argmax_{e' \in d} SIM(e, e')\} \tag{9}$$

Note that, if e is not related to any document entity, then we do not have a corresponding e^* for e. Therefore, the unmatched entity $e \in q \setminus d$ will not expand d. Now, we replace the the transformation F with its value in the Eq.6 to obtain the modified document as follows:

$$d_q = d \cup \{e | e \in q \setminus d \wedge \exists e^* \in d : e^* = argmax_{e' \in d} SIM(e, e')\} \tag{10}$$

The length of the modified document $|d_q|$ is calculated as follows:

$$|d_q| = |d| + \sum_{e \in q \setminus d} \#(e^*; d).SIM(e, e^*) \tag{11}$$

Now, the modified document d_q replace the original document model d in any smoothing method inside language models. As a result, the language models for a query q will be estimated according to the modified document d_q instead of d. We believe that the probability estimation will be more accurate and more effective than ordinary language models. We estimate therefore the following probability $P(e|d_q)$ instead of $P(e|d)$.

- **Entity-Centered Documents**. As we mentioned, $P(e|d_q)$ is normally estimated using maximum likelihood. We propose instead to combine two probabilities to estimate $P(e|d_q)$: the maximum likelihood $P_{ml}(e|d_q)$, and another probability that promotes entity-centered documents, denoted as $P_{ecd}(e|d_q)$. $P_{ecd}(e|d_q)$ is the probability of having a linked entity for e inside the document. We suppose that $P_{ml}(e|d_q)$ and $P_{ecd}(e|d_q)$ are conditionally independent, and therefore we estimate $P(e|d_q)$ using the following equation:

$$P(e|d_q) = P_{ml}(e|d_q) \times P_{ecd}(e|d_q) \tag{12}$$

For the probability $P_{ecd}(e|d_q)$, we propose to estimate it following equation shows:

$$P_{ecd}(e|d_q) = \frac{\sum_{e_i \in d_q} \#(e_i; d_q) \times SIM(e, e_i)}{|d_q|} \tag{13}$$

where SIM is the similarity defined in the equation 1.

The maximum likelihood $P_{ml}(e|d_q)$ is estimated using the following equation:

$$P_{ml}(e|d_q) = \frac{\#(e; d_q)}{|d_q|} \tag{14}$$

Finally, if we take the example of Jelinek-Mercer smoothing, we simply write the extended Jelinek-Mercer smoothing as follows:

$$P_{\lambda}(e|d_q) = (1 - \lambda)P(e|d_q) + \lambda P(e|C) \tag{15}$$

We note that the collection related part of the model is not affected, whereas, the document related probability is differently estimated to consider our two goals.

3 Conclusion

We propose, in this paper, a graph-based model for representing documents and queries. First, we identify entities which are mentioned in a query or a document using Wikipedia. Then, we link between those identified entities based on the structure of Wikipedia, i.e. two entities are linked in a document or a query if there is already a link between them in Wikipedia. Finally, we adapt language models for information retrieval in order to match between document and query graphs. The proposed adaption for language model could be easily applied for any smoothing method like: Dirichlet or Jelinek-Mercer (Zhai & Lafferty, 2004).

For future work, we find many campaigns focusing on entity retrieval evaluation, as a result, many test collection are available for testing our proposed approach. We find among them: Cultural Heritage collections CHIC (Petras *et al.*, 2013, 2012), Entity Retrieval track studies entity retrieval in Wikipedia (Vries *et al.*, 2008; Demartini *et al.*, 2010), the TREC Entity track which defines the related entity finding task (Balog *et al.*, 2009, 2012).

References

ALMASRI M., BERRUT C. & CHEVALLET J.-P. (2013). Wikipedia-based semantic query enrichment. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '13, p. 5–8, New York, NY, USA: ACM.

AUDEH B., BEAUNE P. & BEIGBEDER M. (2014). Exploring query reformulation for named entity expansion in information retrieval. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, p. 929–930, New York, NY, USA: ACM.

BALOG K., DE VRIES A. P., SERDYUKOV P., THOMAS P. & WESTERVELD T. (2009). Overview of the trec 2009 entity track. In *TREC 2009 Working Notes*: NIST.

BALOG K., SERDYUKOV P. & DE VRIES A. P. (2012). Overview of the TREC 2011 entity track. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*: NIST.

DALTON J., DIETZ L. & ALLAN J. (2014). Entity query feature expansion using knowledge base links. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, p. 365–374, New York, NY, USA: ACM.

DEMARTINI G., IOFCIU T. & DE VRIES A. P. (2010). Overview of the inex 2009 entity ranking track. In *Proceedings of the Focused Retrieval and Evaluation, and 8th International Conference on Initiative for the Evaluation of XML Retrieval*, INEX'09, p. 254–264, Berlin, Heidelberg: Springer-Verlag.

GUO J., XU G., CHENG X. & LI H. (2009). Named entity recognition in query. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, p. 267–274, New York, NY, USA: ACM.

LIU X., YANG P. & FANG H. (2014). Entexpo: An interactive search system for entity-bearing queries. In M. DE RIJKE, T. KENTER, A. DE VRIES, C. ZHAI, F. DE JONG, K. RADINSKY & K. HOFMANN, Eds., *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, p. 784–788. Springer International Publishing.

PETRAS V., BOGERS T., TOMS E., HALL M., SAVOY J., MALAK P., PAWŁOWSKI A., FERRO N. & MASIERO I. (2013). Cultural heritage in clef (chic) 2013. In P. FORNER, H. MÜLLER, R. PARDES, P. ROSSO & B. STEIN, Eds., *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, p. 192–211. Springer Berlin Heidelberg.

PETRAS V., FERRO N., GÄDE M., ISAAC A., KLEINEBERG M., MASIERO I., NICCHIO M. & STILLER J. (2012). Cultural heritage in clef (chic) overview 2012.

PONTE J. M. & CROFT W. B. (1998). A language modeling approach to information retrieval. SIGIR '98, p. 275–281: ACM.

VRIES A. P., VERCOUSTRE A.-M., THOM J. A., CRASWELL N. & LALMAS M. (2008). Focused access to xml documents. chapter Overview of the INEX 2007 Entity Ranking Track, p. 245–251. Berlin, Heidelberg: Springer-Verlag.

XU Y., DING F. & WANG B. (2008). Entity-based query reformulation using wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, p. 1441–1442, New York, NY, USA: ACM.

ZHAI C. & LAFFERTY J. (2004). A study of smoothing methods for language models applied to information retrieval. **22**(2), 179–214.

ZIGHELNIC L. & KURLAND O. (2008). Query-drift prevention for robust query expansion. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, p. 825–826, New York, NY, USA: ACM.

Vers une recherche sémantique et à base de graphe dans les systèmes d'accès à l'information juridique

Nada Mimouni, Adeline Nazarenko et Sylvie Salotti

LIPN, CNRS (UMR 7030), Université Paris 13
Sorbonne Paris Cité, F-93430 Villetaneuse
Nada.Mimouni, Adeline.Nazarenko, Sylvie.Salotti@lipn.univ-paris13.fr

Résumé :
Ce papier montre que les systèmes d'accès à l'information juridique pourraient être étendus pour permettre une recherche sémantique mais aussi à base de graphes. Le défi est de retrouver des documents non seulement en fonction de leurs descripteurs de contenu, mais aussi sur la base des relations intertextuelles qu'ils entretiennent. Le papier présente une formalisation logique pour décrire les collections documentaires et un langage de requêtes simple basé sur les graphes qui est défini pour répondre aux besoins du projet Légilocal.
Mots-clés : Recherche d'information sémantique, analyse des besoins, langage de requêtes.

1 Introduction

La modélisation de l'intertextualité forme un enjeu dans le domaine juridique où les nouveaux documents (décisions administratives, les jugements, les provisions) reposent sur ceux déjà existants dont ils modifient, annulent, confirment ou appliquent dans un nouveau contexte. L'intertextualité a été identifiée comme une source majeure de complexité de la documentation juridique (Bourcier, 2011). Elle a été analysée au niveau global du réseau documentaire mais pas au niveau local qui est plus pertinent dans le cas de production de textes législatifs (*legal drafting*) ou l'analyse des cas.

La majorité des systèmes d'accès à l'information juridique permettent aux praticiens de retrouver des documents, mais pas sur des critères intertextuels. Ce phénomène d'intertextualité est apparu comme une limitation importante dans le contexte du projet Légilocal [1] où les agents administratifs sont amenés de façon quotidienne à analyser et à produire des documents juridiques (Amardeilh *et al.*, 2013). Lors de la rédaction d'une ordonnance, les secrétaires de mairies doivent généralement identifier la législation à laquelle il faut se référer, les anciens actes publiés sur le même sujet et en particulier ceux qui ont fait l'objet d'un recours.

1. Ce travail a été partiellement financé par le projet LEGILOCAL (FUI-9, 2010-2013) et par le Labex EFL (ANR-10-LABX-0083).

Ce papier présente une formalisation logique qui a été définie pour décrire et interroger les collections documentaires sur des bases intertextuelles. Il ouvre la voie vers de nouvelles fonctionnalités de recherche prenant en compte les relations intertextuelles. L'opérationnalisation a été faite par deux approches : une approche de classification conceptuelle utilisant l'analyse formelle et relationnelle de concept (Mimouni *et al.*, 2015a) et une approche utilisant les technologies du web sémantique (Mimouni *et al.*, 2015b).

La suite du papier est organisée comme suit. La section 2 définit l'objectif de cette proposition vis à vis des travaux précédents dans la recherche d'information et l'analyse des liens. Les sections 3 et 4 présentent la collection Légilocal comme un réseau sémantique de documents et le type du langage de requêtes qui peut être utilisé pour l'interroger. La section 5 montre comment ce langage de requêtes peut être utilisé pour exprimer un échantillon de requêtes exprimées par des praticiens juridiques et collectées dans le cadre du projet Légilocal. La section 6 discute les limitations du langage et les extensions possibles.

2 L'intertextualité dans la recherche d'information juridique

L'« intertextualité », dans un usage restreint, est traditionnellement définie comme « une relation de co-présence entre deux ou plusieurs textes, à savoir, le plus souvent, par la présence effective d'un texte dans un autre » (Genette, 1982, p.8). Les citations - forme explicite de l'intertextualité - sont importantes à prendre en compte lorsque les textes cibles contribuent à l'interprétation des textes sources, comme il est souvent le cas pour les textes juridiques (Bhatia, 1998).

Plusieurs travaux ont reconnu l'importance de l'intertextualité des sources juridiques qui représente un facteur majeur de complexité de la documentation (Bourcier, 2011). L'analyse de réseaux a été considérée comme un moyen puissant pour modéliser les collections juridiques (Fowler *et al.*, 2007; Romain *et al.*, 2011; Winkels & de Ruyter, 2011). Cependant, comme dans l'analyse de citations et de réseaux sociaux (Rubin, 2010), l'accent a été mis sur le niveau du réseau afin d'identifier les sous-collections les plus fortement connectées ou les sources de loi les plus influentes. Moins d'attention a été accordée à l'analyse détaillée de l'intertextualité et la sémantique des liens intertextuels.

En recherche d'information, les liens intertextuels ou les citations sont généralement modélisés comme des métadonnées associées aux documents qui sont présentés aux utilisateurs. Ces derniers peuvent naviguer d'un document aux sources qu'il cite, puis à partir de ces sources vers les documents auxquels ils se réfèrent, et ainsi de suite. Ceci représente la façon hypertextuelle commune de manipulation de l'intertextualité, où l'on se perd rapidement dans l'hyperespace (Conklin, 1987).

L'intertextualité a été aussi utilisée pour le tri des documents, comme dans Brin & Page (1998), ce qui est moins pertinent dans le domaine juridique, où l'exhaustivité de la

recherche est plus importante que le classement des résultats.

Les recherches récentes dans l'analyse socio-sémantique exploitent à la fois la topologie des réseaux et la sémantique de leurs noeuds et liens (par ex. Cointet & Roth (2009)). Cela ouvre la voie à de nouvelles fonctionnalités de recherche (par ex. le *Graph Search* de Facebook). En se basant sur ces premières expériences, nous défendons l'idée que les systèmes d'accès à l'information juridique peuvent aller plus loin et exploiter sémantiquement l'intertextualité, c.à.d. comme un critère de recherche.

Dans ce qui suit, nous montrons l'avantage que les praticiens du droit peuvent tirer d'un langage de requêtes relationnelles lors de la rédaction des actes administratifs.

3 La collection Légilocal comme un réseau sémantique de documents

Nous considérons une collection juridique comme étant l'ensemble des documents reliés par des liens intertextuels de types différents. La collection de documents Légilocal consiste en un nombre croissant de documents produits ou cités par le réseau des secrétaires de mairies dans Légilocal qui collaborent pour l'élaboration des actes administratifs.

3.1 Description de la collection

Nous considérons tout fragment de document qui peut être mis à jour ou retourné indépendamment de son document comme une unité documentaire. Dans Légilocal, tout document ou article appartenant à un document juridique est une unité documentaire et chaque unité a un identifiant unique. Dans ce qui suit, d_i se réfère à l'unité documentaire i.

La collection est composée de différents types de documents : actes administratifs utilisés comme exemples positifs ou négatifs dans le processus de rédaction, divers textes législatifs issus de juridictions supérieures et utilisées comme référence, ainsi que des documents éditoriaux (exemples et directives). Tout document d_i possède un type unique j : $Type(d_i, t_j)$ (dans ce qui suit, t_j se réfère au type j).

Les documents sont annotés. Nous nous focalisons ici sur les annotations sémantiques, qui sont les mots-clés du texte ou les tags des utilisateurs associés aux documents à des fins de recherche. Tout document est associé à un nombre quelconque d'attributs. Dans ce qui suit, $Att(d_i, s_k)$ indique que le descripteur s_k est attaché au document d_i.

Les documents (ou unités documentaires) sont reliés les uns aux autres par différents types de liens dont la sémantique dépend des types des documents reliés et de la valeur du lien lui-même. Ces liens sont orientés : $Rel(d_i, r_l, d_{i'})$ indique que d_i est la source d'un lien r_l dont la cible est $d_{i'}$ [2].

2. Nous simplifions cette représentation de plusieurs manières : i) les types et les descripteurs sémantiques sont en fait organisés en hiérarchies, ii) des relations ternaires existent aussi, par exemple lorsqu'un

Types	Descriptifs
Décision	décision
ArrêtéMun	arrêté municipal
ArrêtCcass	Arrêt de Cour de cassation
ArrêtCappel	Arrêt de Cour d'appel
ArticleCode	Article de code
Relations	**Descriptifs**
application	un texte législatif en *applique* un autre ou une décision *applique* une autre décision ou un texte législatif
décision	un jugement fait une *décision* sur un jugement précédent
annulation	la décision ci-dessus est une *annulation*
confirmation	la décision ci-dessus est une *confirmation*
composition	un document *se compose d'*articles
Descripteurs	**Equivalents terminologiques**
cheminR	« chemins rural »
véhiculeAMoteur	« véhicule à moteur »
Identifiants	**Référents**
Code_{Env}	Code de l'environnement
$\text{Code}_{CV}\text{\_Article}_{1382}$	Article 1382 du Code civil
ArrêtCcass_{A}	Arrêté A de la Cour de cassation
ArrêtCappel_{X}	Arrêté X de la Cour d'appel

TABLE 1 – Vocabulaire utilisé pour la formation de la collection Légilocal et des requêtes associées

Le tableau 1 présente une sélection du vocabulaire (types de documents et identifiants, descripteurs sémantiques et relations) utilisé pour la description de la collection Légilocal.

3.2 Formalisation de la collection

À partir de cette analyse, une collection documentaire C peut être modélisée comme un graphe orienté, étiqueté et attribué $C = \mathcal{G}(D, R, A)$ où

- les noeuds sont des unités documentaires de D ;
- les unités documentaires sont décrites par des attributs, types de T et descripteurs sémantiques de S ($A = T \cup S, T \cap S = \emptyset$) ;
- les arcs sont des relations binaires typées et orientées, avec des types appartenant à R.

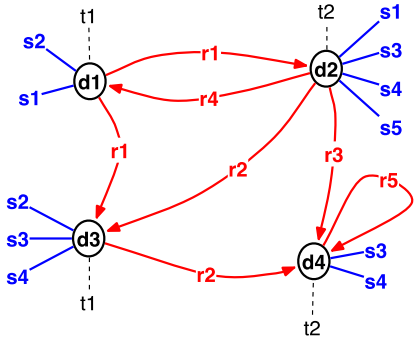
document d_1 indique qu'un document d_2 est modifié et remplacé par un document d_3 et iii) nous ne considérons pas la distinction œuvre *vs.* expression (Rubin, 2010; Sartor *et al.*, 2011).

$graph_{coll} \rightarrow predicate_c$ ['$\wedge$ ' $predicate_c$]*
$predicate_c \rightarrow$ 'Type' '('id_{doc}','id_{type}')' | 'Att' '('id_{doc}',' id_{sem}')' | 'Rel' '('id_{doc}',' id_{rel}',' id_{doc}')'
$id_{doc} \rightarrow$ 'd_1' | 'd_2' | ...
$id_{type} \rightarrow$ 't_1' | 't_2' | ...
$id_{sem} \rightarrow$ 's_1' | 's_2' | ...
$id_{rel} \rightarrow$ 'r_1' | 'r_2' | ...
where ($\forall$ i, j, k, l) ($d_i \in$ D, $s_j \in S$, $t_k \in T$ and $r_l \in R$).

FIGURE 1 – Langage modélisant les collections documentaires. Les éléments du vocabulaire terminal sont notés entre guillemets simples (ex. '('), les non-terminaux sont en italiques (ex. *prédicat*) et les métasymboles utilisés sont la flèche de réécriture ($\rightarrow$), les crochets pour former les groupes ([ ]), la barre d'alternative (|) et l'étoile de Kleene pour marquer la répétition de l'élément ou du groupe précédent pour un nombre quelconque d'occurrences (*).

Notons qu'il n'existe aucune contrainte sur le nombre de noeuds, attributs et liens dans le graphe ni sur la combinaison des attributs et liens pour une unité documentaire donnée.

Une telle collection est décrite par une formule du langage présenté dans la figure 1. La figure 2 montre le graphe d'un exemple de collection associé avec sa formule.



$$Type(d_1, t_1) \wedge Att(d_1, s_1) \wedge Att(d_1, s_2) \wedge Type(d_2, t_2) \wedge Att(d_2, s_1) \wedge Att(d_2, s_3) \wedge Att(d_2, s_4) \wedge Att(d_2, s_5) \wedge Type(d_3, t_1) \wedge Att(d_3, s_2) \wedge Att(d_3, s_3) \wedge Att(d_3, s_4) \wedge Type(d_4, t_2) \wedge Att(d_4, s_3) \wedge Att(d_4, s_54) \wedge Rel(d_1, r_1, d_2) \wedge Rel(d_2, r_4, d_1) \wedge Rel(d_1, r_1, d_3) \wedge Rel(d_2, r_2, d_3) \wedge Rel(d_2, r_3, d_4) \wedge Rel(d_3, r_2, d_4) \wedge Rel(d_4, r_5, d_4)$$

FIGURE 2 – Exemple de graphe modélisant une collection documentaire comportant 4 unités documentaires. Les attributs et relations partagés par plusieurs documents sont représentés en double. Les unités documentaires sont représentées par des cercles. Les relations sont notées comme des flèches. Les attributs sont reliés aux documents par des traits pleins (descripteurs sémantiques) ou pointillés (types de documents).

‘graph$_{query}$’ → [*focus* ‘ :’] ? *graph$_q$* [‘with’ *constraint* [‘∧’ *constraint*]*] ?
focus → ‘(’ *variable* [‘,’ *variable*]* ‘)’
graph$_q$ → *predicate$_q$* [‘∧’ *predicate$_q$*]*
predicate$_q$ → ‘Att’ ‘(’ *document* ‘,’ *attribute* ‘)’ | ‘Type’ ‘(’ *document* ‘,’ *type*‘)’ | ‘Rel’ ‘(’ *document* ‘,’ *relation* ‘,’ *document* ‘)’
document → *id$_{doc}$* | *var$_{doc}$*
attribute → *id$_{sem}$* | *vat$_{sem}$*
type → *id$_{type}$* | *var$_{type}$*
relation → *id$_{rel}$* | *var$_{rel}$*
id$_{doc}$ → ‘d$_1$’ | ‘d$_2$’ | ‘d$_3$’ | . . .
id$_{sem}$ → ‘s$_1$’ | ‘s$_2$’ | ‘s$_3$’ | . . .
id$_{type}$ → ‘t$_1$’ | ‘t$_2$’ | ‘t$_3$’ | . . .
id$_{rel}$ → ‘r$_1$’ | ‘r$_2$’ | ‘r$_3$’ | . . .
constraint → *variable* ‘≠’ *variable*
variable → *var$_{doc}$* | *var$_{sem}$* | *var$_{type}$* | *var$_{rel}$*
where *var$_{doc}$* ∈ D, *var$_{sem}$* ∈ S, *var$_{type}$* ∈ T, *var$_{rel}$* ∈ R
and (∀ i, j, k, l) (d$_i$ ∈ D ∧ s$_j$ ∈ S ∧ t$_k$ ∈ T ∧ r$_l$ ∈ R)

FIGURE 3 – Langage de requêtes. La convention de notation est la même que pour la figure 1. Le méta-symbole *?* indique que l’élément ou le groupe précédent se produit au plus une fois.

4 Requêtes relationnelles et structurées

Modéliser les collections documentaires comme des graphes nous amène à considérer l’interrogation avec des graphes (Khan *et al.*, 2012) comme une approche de recherche d’information où les requêtes se formalisent elles même sous forme de graphes. Un graphe de requête est similaire à celui de la collection mais qui peut contenir :

– des variables à la place des identifiants des documents, attributs, types et relations,
– des contraintes d’inégalité sur ces variables,
– une cible pour restreindre la réponse à un sous ensemble des variables de la requête.

Un graphe requête est donc décrit par une formule du langage donné par la grammaire de la figure 3.

L’appariement entre les requêtes et les documents revient à instancier le graphe de requête sur le graphe de la collection. Une réponse est :

– un ensemble contenant tous les sous-graphes du graphe de la collection qui instancient le graphe requête s’il ne possède pas une cible explicite,
– un ensemble contenant tous les tuples d’identifiants qui instancient la cible de la requête si elle possède un,

– un ensemble vide si le graphe requête ne peut être instancié.

Prenons quelques exemples de requêtes (avec et sans cible et contraintes) ainsi que les réponses produites par leur appariement sur l'échantillon de la collection de la figure 2 :

1. $Att(x, s_1) \wedge Rel(x, r_1, d_2)$
   Trouver tous les sous-graphes composés d'un document décrit par s_1 et ayant pour cible d_2 par la relation r_1.
   Résultat : $\{Att(d_1, s_1) \wedge Rel(d_1, r_1, d_2)\}$ (1 graphe)
2. $Rel(x, y, x)$
   Trouver tous les documents en relation avec eux-même.
   Résultat : $\{Rel(d_4, r_5, d_4)\}$ (1 graphe)
3. $(y) : Rel(x, y, x)$
   Trouver tous les types de relation reliant un document à lui-même.
   Résultat : $\{r_5\}$ (1 relation)
4. $(x, y) : Rel(x, y, x)$
   Trouver tous les couples composés d'un document lié à lui-même et du type de la relation.
   Résultat : $\{(d_4, r_5)\}$ (1 couple composé d'un document et d'une relation)
5. $(x, y) : Att(x, z) \wedge Rel(x, r_1, y) \wedge Att(y, z)$
   Trouver tous les couples de documents décrits par un même descripteur sémantique et tels que le second est la cible du premier par la relation r_1.
   Résultat : $\{(d_1, d_2), (d_1, d_3)\}$ (2 couples)
6. $Att(x, s_2) \wedge Att(x, y)$ avec $y \neq s_2 \wedge y \in S$
   Trouver tous les sous-graphes composés d'un document décrit par s_2 et un autre descripteur sémantique différent.
   Résultat : $\{Att(d_1, s_1) \wedge Att(d_1, s_2), Att(d_3, s_2) \wedge Att(d_3, s_3), Att(d_3, s_2) \wedge Att(d_3, s_4)\}$ (3 graphes)
7. $Type(x, y) \wedge Type(x, z)$ avec $y \neq z$
   Trouver les documents de deux types différents.
   Résultat : $\emptyset$

5 L'intertextualité dans les requêtes des praticiens

Le langage de requête ci-dessus offre une manière homogène et naturelle pour exprimer un large éventail de requêtes des praticiens, qui combinent souvent le contenu et les critères intertextuels. Les exemples de requêtes suivantes ont été toutes recueillies dans le cadre de l'analyse des besoins du projet Légilocal. Nous montrons comment elles peuvent

être formalisées utilisant le langage de requêtes ci-dessus, en laissant de côté les questions relatives à la formulation des requêtes et leur réponses. Cette formalisation suppose que les documents sont correctement analysés, leurs types sont identifiés, ils sont annotés avec des descripteurs sémantiques et les liens intertextuels sont eux-mêmes identifiés et sémantiquement typés.

1. "Quelles sont les décisions de jurisprudence qui citent l'article 1382 du code civil ?"
   $(x) : Type(x, Decision) \wedge Rel(x, application, Code_{CV}\_Article_{1382})$
   Le terme générique "cite" est interprété comme une relation d'application à cause des types des documents reliés, une décision de jurisprudence et un texte législatif.
2. "Quels sont les articles du code de l'environnement qui parlent de véhicules à moteurs ?"
   $Type(x, ArticleCode) \wedge Rel(x, composition, Code_{Env}) \wedge Att(x, vehiculeAMoteur)$
3. "Je voudrais la décision qui fait l'objet de l'arrêt A de la Cour de cassation."
   $(x) : Type(x, Decision) \wedge Rel(ArretCcass_A, decision, x)$
4. "Je cherche les décisions qui ont été annulées par la Cour de cassation."
   $(x) : Type(x, Decision) \wedge Rel(y, annulation, x) \wedge Type(y, ArretCcass)$
   $Type(x, Decision) \wedge Rel(y, decision, x) \wedge Type(y, ArretCcass)$
   Ces deux formules de requêtes ne diffèrent que par leurs cibles : dans la première, la requête devrait avoir comme réponse une ou plusieurs décisions tandis que la deuxième aurait comme réponse une liste de graphes qui instancient le graphe requête (voir figure 4).

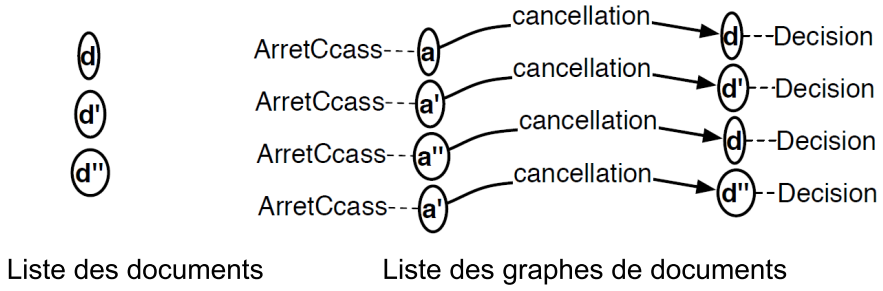


FIGURE 4 – Exemples de réponses associées à la requête 4 : une liste de documents (1ère formule) ou une liste de graphes de documents (2ème formule).

5. "Je voudrais savoir si cet arrêt de la cour d'appel a fait lui-même objet d'un recours."
   $Rel(x, decision, ArretCappel) \wedge Type(x, ArretCcass)$ [3]
6. "Je cherche des arrêtés municipaux concernant les chemins ruraux qui ont fait l'objet d'un recours et ont été annulés par une décision de jurisprudence."
   $(x) : Type(x, ArreteMun) \wedge Att(x, cheminR) \wedge Rel(y, annulation, x)$

3. Le type du jugement final ($Att(x, ArretCcass)$) dépend de la procédure d'appel suivie.

7. Quels sont les articles de code qui ont été confirmés et qui sont cités par les arrêtés municipaux parlant de chemins ruraux ?"
   $(x) : Type(x, ArticleCode) \wedge Type(y, ArreteMun) \wedge Att(y, cheminR) \wedge Rel(y, application, x) \wedge Rel(z, confirmation, y)$
8. "Je souhaite savoir si les textes visés par des arrêtés municipaux parlant de chemins ruraux sont aussi cités par ceux concernant les véhicules à moteurs."
   $Type(x, ArreteMun) \wedge Att(x, cheminR) \wedge Rel(x, application, y) \wedge Rel(z, application, y) \wedge Type(z, ArreteMun) \wedge Att(z, vehiculeAMoteur)$
9. "Quels sont les arrêtés municipaux qui ont fait l'objet de deux recours ?"
   $(x) : Type(x, ArreteMun) \wedge Rel(y, decision, x) \wedge Rel(z, decision, x)$ with $y \neq z$

6 Discussion

Pour des fins de démonstration, nous gardons le langage de requête aussi simple que possible. Nous nous sommes concentrées sur l'intertextualité, mais le langage de requête ci-dessus doit naturellement être étendu pour interroger les documents avec leurs métadonnées (par ex. la date de publication ou l'auteur), les documents sources (les œuvres) doivent être différenciés de leurs versions (expressions), tel que proposé par Sartor *et al.* (2011).

Le langage de requêtes ci-dessus permet de traiter l'intertextualité mais présente des limites.

Quantification Les requêtes en langage naturel impliquent des hypothèses de (non-)unicité qui ne sont pas exprimables dans le langage proposé. Par exemple, les variantes de requêtes suivantes sont considérées comme équivalentes dans notre langage de requête, où les variables sont quantifiées de manière existentielle : « Quels sont les jugements qui confirment une/des/plusieurs décision(s) … ». Même si la quantification universelle permettrait d'exprimer des requêtes telles que « Y a t-il un article de code cité par tous les arrêtés portant sur les chemins ruraux ? », nous avons choisi de ne pas l'inclure dans un premier temps, car elle est difficile à maîtriser pour les utilisateurs et qu'elle n'apparaissait pas dans les requêtes recueillies.

Négation et disjonction Pour préserver la simplicité du langage pour les utilisateurs, nous avons choisi de ne pas inclure la négation ou la disjonction des opérateurs dans la spécification du langage de requête, ce qui est une limitation en ce qui concerne les besoins des praticiens. Par exemple, la requête « Quelles sont les décisions antérieures à la décision D ? » doit être formulée de la manière suivante :

$$(x) : Type(x, Decision) \wedge (Rel(decision_D, decision, x) \vee (Rel(decision_D, decision, y) \wedge Rel(y, decision, x)))$$

pour prendre en compte différentes longueurs de chaînes de décision. Aussi, sans opérateur de négation, une requête comme « Quels sont les articles qui ne sont pas annulés ? » ne peut être formalisée que comme « Quels sont les articles qui ont été confirmés ? », qui est plus restrictive.

Cible de requête Il est souvent difficile d'identifier si une requête en langage naturel est ciblée ou non. Même si on est habitué à avoir des listes de documents, nous nous attendons à ce que les utilisateurs spécialisés apprécient un large éventail de types de réponses. Les graphes réponses donnent plus de contexte et peuvent être affinés grâce à une interface interactive. La différence ne réside pas dans la mise en correspondance du graphe de la requête et de la collection, mais dans la présentation des résultats.

Opérateur de comptage Jusqu'à présent, nous n'avons recueilli aucune requête nécessitant un opérateur de comptage, mais ce point doit être étudié davantage.

Topologie de graphe Nous n'avons mis aucune contrainte sur la taille des graphes de requêtes ni sur la présence de cycles. Même si les exemples ci-dessus de graphes de requêtes sont simples, nous nous attendons à ce que les utilisateurs spécialisés entrent progressivement des requêtes plus complexes.

7 Conclusion

L'analyse des besoins dans le projet Légilocal a révélé les limitations des systèmes existants d'accès à l'information juridique, qui permettent aux utilisateurs de retrouver des documents en fonction de leurs métadonnées et les descripteurs de contenu mais pas en fonction des relations intertextuelles qui sont néanmoins critiques dans l'analyse juridique.

En outre, l'état de recherche et des technologies permet aujourd'hui de développer des approches socio-sémantiques de recherche (Cointet & Roth, 2009) qui permettent de rechercher dans les grands graphes attribués exploitant à la fois les attributs des noeuds et la structure du réseau. Nous soutenons l'idée qu'une approche similaire peut être adoptée pour la recherche d'information juridique.

Les collections juridiques peuvent être modélisées comme des réseaux sémantiques de documents, où les noeuds (les documents) sont associés à des attributs sémantiques et sont reliés les uns aux autres par divers types de liens sémantiques.

Ce papier montre l'avantage qui peut être tiré d'un simple langage de graphe de requêtes et ouvre la voie vers une nouvelle forme de recherche sémantico-relationnelle dans les sources juridiques.

Beaucoup de travail reste à faire. Nous considérons que les premiers outils de démonstration doivent être très simples et seront enrichis progressivement par de nouvelles fonctionnalités lorsque les utilisateurs prennent l'habitude des premiers tests. Toutefois, exploiter un langage de graphe de requêtes, aussi simple qu'il soit, nécessite une interface utilisateur adéquate. Parmi les différents types de modes d'interrogation, ceux basés sur un formulaire ou sur la technique "*fill-in-the-blank*" sont probablement les plus commodes, mais cela reste à évaluer. L'approche proposée doit être testée sur une collection de taille croissante au sein du projet Légilocal. L'interrogation est actuellement mise en œuvre avec SPARQL et des outils sont développés en parallèle pour annoter les documents de la collection et construire le réseau sémantique des documents Légilocal.

Références

AMARDEILH F., BOURCIER D., CHERFI H., DUBAIL C., GARNIER A., GUILLEMIN-LANNE S., MIMOUNI N., NAZARENKO A., PAUL ÈVE., SALOTTI S., SEIZOU M., SZULMAN S. & ZARGAYOUNA H. (2013). The légilocal project : the local law simply shared. In K. D. ASHLEY, Ed., *Legal Knowledge and Information Systems - JURIX 2013 : The Twenty-Sixth Annual Conference*, volume 259, p. 11–14, University of Bologna, Italy : IOS Press.

BHATIA V. K. (1998). Intertextuality in legal discourse. *JALT Journal Online*, (1).

BOURCIER D. (2011). Sciences juridiques et complexité. un nouveau modèle d'analyse. *Droit et Cultures*, **61**(1), 37–53.

BRIN S. & PAGE L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web (WWW7)*, p. 107–117, Amsterdam, The Netherlands : Elsevier Science Publishers B. V.

COINTET J. & ROTH C. (2009). Socio-semantic dynamics in a blog network. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 4, p. 114–121.

CONKLIN J. (1987). Hypertext : An introduction and survey. *IEEE Computer*, **20**(9), 17–41.

FOWLER J. H., JOHNSON T. R., SPRIGGS J. F., JEON S. & WAHLBECK P. J. (2007). Network analysis and the law : Measuring the legal importance of precedents at the u.s. supreme court. *Political Analysis*, **15**, 324–346.

GENETTE G. (1982). *Palimpsestes*. Poétique. Le Seuil.

KHAN A., WU Y. & YAN X. (2012). Emerging graph queries in linked data. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, p. 1218–1221, USA.

MIMOUNI N., NAZARENKO A. & SALOTTI S. (2015a). A conceptual approach for relational IR : application to legal collections. In J. BAIXERIES, C. SACAREA & M. OJEDA-ACIEGO, Eds., *Formal Concept Analysis - 13th International Conference, ICFCA 2015, Nerja, Spain, June 23-26, 2015, Proceedings*, volume 9113 of *Lecture Notes in Computer Science*, p. 303–318 : Springer.

MIMOUNI N., NAZARENKO A. & SALOTTI S. (2015b). Une ontologie documentaire pour l'accès aux contenus juridiques. In *26es Journées francophones d'Ingénierie des Connaissances (IC), Vers le traitement de la masse de données disponibles sur le web, 29 Juin - 3 Juillet, 2015*,

Rennes - France.

ROMAIN, MAZZEGA P. & BOURCIER D. (2011). A network approach to the french system of legal codes- part i : Analysis of a dense network. *Journal of Artificial Intelligence and Law*, **19**, 333–355.

RUBIN R. (2010). *Foundations of Library and Information Science*. Neal-Schuman Publishers.

SARTOR G., PALMIRANI M., FRANCESCONI E. & BIASIOTTI M. A. (2011). *Law, Governance and Technology : Legislative Xml for the Semantic Web : Principles, Models, Standards for Document Management*. Law, Governance and Technology Series, 4. Springer London, Limited.

WINKELS R. & DE RUYTER J. (2011). Survival of the fittest : Network analysis of dutch supreme court cases. In *AICOL*, p. 106–115.

Extraction automatique de relations sémantiques définies dans une ontologie

Albert Royer, Christian Sallaberry, Annig Le Parc-Lacayrelle, Marie-Noëlle Bessagnet

LIUPPA Laboratoire LIUPPA, France
albert.royer@univ-pau.fr
christian.sallaberry@univ-pau.fr
annig-lacayrelle@univ-pau.fr
marie-noelle.bessagnet@univ-pau.fr

Résumé : Cette contribution se situe dans le domaine de la recherche d'information sémantique et s'intéresse plus particulièrement à la phase d'annotation. Nous proposons une méthode qui permet d'annoter automatiquement des documents textuels sur la base de concepts et de relations sémantiques modélisés dans une ontologie. Cette méthode est générique car elle est capable d'exploiter le contenu sémantique de toute ontologie. Elle a été mise en œuvre sur la plateforme GATE [1].

Abstract : This contribution is in the field of semantic information retrieval from textual documents and more particularly the annotation stage. We propose an automatic annotation method based on concepts and semantic relationships modeled in an ontology. This method is generic as it is able to exploit the semantic content of any ontology. It has been implemented on the GATE platform.

Mots-clés : recherche d'information sémantique, annotation automatique de concept, annotation automatique de relation sémantique, ontologie

1 Introduction

Pour faire face à l'augmentation exponentielle de données désormais disponibles dans des formats numériques, les modèles et technologies supportant les moteurs de recherche ont intégré de nombreuses améliorations (Buscaldi *et al.*, 2013). Toutefois, ces propositions restent souvent limitées par l'usage de mots-clés, qui contraste avec l'idée de recherche « sémantique » où les descripteurs d'une collection de documents ou d'un besoin d'information sont des entités sémantiques (représentant le sens d'un mot ou d'un syntagme). Ce paradigme de recherche, appelé recherche d'information sémantique (RIS), exploite généralement des ressources telles que des thésaurus ou des ontologies dans les phases d'indexation et de recherche.

La difficulté de production de telles ressources sémantiques est réelle. Elle combine souvent des techniques automatiques et l'intervention d'experts de domaines spécifiques. Dans ce travail, nous considérons que ces ressources existent et nos propositions se focalisent sur leur exploitation plutôt que leur construction. Il n'en demeure pas moins que les processus d'annotation automatique de concepts et de relations, en vue de l'extraction de descripteurs sémantiques, sont dépendants de la qualité de ces ressources. Ainsi, les résultats obtenus pour des approches de RIS sont souvent mitigés lorsqu'on les compare à la recherche d'information (RI) classique. Ils sont cependant encourageants pour des domaines spécifiques qui peuvent être décrits plus

1. https://gate.ac.uk/

facilement (Kiryakov *et al.*, 2004). Nous pouvons citer des expériences positives spécifiques au domaine médical (Abasolo & Gomez, 2000; Trieschnigg *et al.*, 2009), par exemple. RIS et RI sont également combinées dans certaines approches (Buscaldi & Zargayouna, 2013; Kara *et al.*, 2012).

Notre proposition se situe dans le contexte de la RIS. Nous avons conçu et mis en œuvre le prototype de RIS ThemaStream (Buscaldi *et al.*, 2013) qui exploite une ontologie de domaine dédiée aux plantes. Comme pour d'autres prototypes (Kara *et al.*, 2012), le processus d'annotation des relations sémantiques s'appuie sur des algorithmes dépendants du domaine et ne traite pas les problèmes d'ambiguïté. Notre contribution est une nouvelle démarche automatique d'annotation de relations sémantiques dans des textes, indépendante du domaine applicatif. Nous proposons un algorithme de recherche de triplets « domaine/relation/codomaine » qui prend la relation comme point de départ et non pas les concepts, comme dans la majorité des approches.

L'originalité de cette proposition est sa généricité. En effet, l'algorithme d'annotation de relations sémantiques est indépendant du domaine. Il exploite les concepts et les relations sémantiques de toute ontologie.

Dans cet article, nous exposons notre processus de reconnaissance, dans des documents textuels, de relations sémantiques préalablement définies dans une ontologie. Après cette partie introductive, une deuxième partie présente succinctement des travaux de recherche en lien avec notre contribution. Une troisième partie décrit le modèle de concept et de relation sur lequel reposent nos propositions. Elle présente deux algorithmes dédiés au marquage de relations sémantiques dans des textes. Une quatrième partie illustre ces propositions à travers des exemples d'ontologie, de texte à analyser et de recherche d'information. Nous terminons par une conclusion et des perspectives.

2 Travaux connexes

L'exploitation de la sémantique est au centre des travaux de chercheurs de différents domaines : la représentation et la gestion des connaissances, le web sémantique (WS) ou web de données et la RI. Quel que soit le domaine, un ensemble de connaissances peut être modélisé sous la forme d'une ontologie exploitée localement ou partagée via le WS (Linked Data ou Linked Open Data).

La RI est la tâche de recherche de documents, au sein d'une collection, satisfaisant un besoin d'information (Manning *et al.*, 2008). Les premières propositions relatives à la RIS, au-delà des modèles de recherche basés sur les seuls mots-clés, visent l'exploitation de vocabulaires (WordNet [2], par exemple) qui associent du sens à des termes (Kara *et al.*, 2012). Ces vocabulaires permettent par exemple l'expansion d'index et de requêtes dans des processus de RI. De manière générale, la RIS exploite des descripteurs sémantiques de documents pour répondre à un besoin d'information.

La première difficulté est donc l'annotation sémantique en amont de l'étape de RIS proprement dite. L'annotation de textes fixe l'interprétation d'un document en lui associant une sémantique formelle et explicite (Kiryakov *et al.*, 2004). Elle consiste à associer des descripteurs relatifs au contenu, à la structure, ou encore au contexte des documents textuels.

2. `http://wordnet.princeton.edu/`

(Kara *et al.*, 2012; Fernandez *et al.*, 2011) distinguent les approches d'annotation sémantique automatique basées sur l'exploitation (1) de la langue, (2) de modèles statistiques et (3) d'ontologies.

La première catégorie d'approches est basée sur le traitement automatique de la langue et exploite des patrons linguistiques définis «manuellement» par des experts. Elle nécessite des ressources importantes en termes de capacité de traitement mais aboutit généralement à des taux de rappel et de précision satisfaisants.

La seconde catégorie est basée sur des techniques d'apprentissage. Les approches de cette catégorie peuvent être supervisées. Dans ce cas, les règles d'annotation sont déduites automatiquement à partir de l'analyse d'un échantillon de documents manuellement associés à des classes (catégories) par des experts ; citons SVM (Support Vector Machine) et K-NN (k-Nearest Neighbors). Elles peuvent être aussi non supervisées et associées à des ressources externes, auquel cas les classes sont définies à partir de ces ressources et les règles d'annotation déduites automatiquement de l'analyse d'un échantillon de documents. On parle d'approches ressource-dépendantes : par exemple, la méthode ESA (Explicit Semantic Analysis) est associée à des bases de connaissances telles que Wikipedia. Enfin, ces méthodes peuvent être non supervisées et ne requérir aucune ressource externe. On parle d'approches corpus-dépendantes : les classes sont déduites de l'analyse d'un échantillon de documents. Citons les méthodes LDA (Latent Dirichlet Allocation), LSA (Latent Semantic Analysis), pLSA (probabilistic LSA) et LSI (Latent Semantic Indexing). Moins exigeantes en termes de capacité de traitement, elles aboutissent généralement à des taux de rappel et de précision moins importants que ceux de la première catégorie.

La troisième catégorie d'approches, quant à elle, s'appuie sur l'exploitation de bases de connaissances ontologiques (Ontologie Based Information Extraction - OBIE). L'annotation est guidée par les connaissances modélisées dans l'ontologie : classes, concepts, relations (Wimalasuriya & Dou, 2010; Nebhi, 2012). Au-delà de la simple extraction d'information, il s'agit d'enrichir l'annotation d'un document par des liens vers des connaissances supplémentaires décrites dans une ontologie locale (Wang & Stewart, 2015) ou partagée via le WS (Nebhi, 2012). L'architecture générale d'un système OBIE est présentée dans (Wang & Stewart, 2015). L'ontologie y est considérée comme une ressource qui peut-être soit fournie au système en entrée, soit construite et mise à jour par le système.

Nos travaux se situent dans cette troisième catégorie. Comme la plupart des systèmes OBIE (SOBA, KIM, PANKOW cités dans (Wang & Stewart, 2015)) nous avons adopté l'approche basée sur une ontologie existante fournie en entrée.

La table 1 liste de nombreux prototypes de RIS qui s'appuient sur l'exploitation de bases de connaissances ontologiques. Ces prototypes utilisent des ontologies ou le WS pour annoter des concepts (C) uniquement ou des concepts et des relations (C et R).

La plupart de ces systèmes sont liés à un domaine, comme celui proposé par (Kara *et al.*, 2012) ou encore dans nos précédents travaux relatifs à ThemaStream (Buscaldi *et al.*, 2013). Notre contribution propose une nouvelle démarche automatique d'annotation de relations sémantiques dans des textes. Cette approche est générique et ouvre ainsi la possibilité d'exploiter les concepts et les relations sémantiques définies dans toute ontologie. Ce travail va dans le sens des préconisations de (Lee *et al.*, 2014) qui mettent en exergue l'importance de l'exploitation des relations sémantiques entre concepts dans le processus de RIS.

Prototype de RIS	Domaine	Annotation
TextViz (Dudognon *et al.*, 2010)	Ontologie de domaine (Mécanique)	C
(Fernandez *et al.*, 2011)	WS Ontologie de domaine (Football)	C C
(Kara *et al.*, 2012)	Ontologie de domaine (Football)	C et R
ThemaStream (Buscaldi *et al.*, 2013)	Ontologie de domaine (Plantes)	C et R
YaSemIR (Buscaldi & Zargayouna, 2013)	Ontologie (domaine indifférent)	C
Broccoli (Bast *et al.*, 2014)	WS (Plantes sur Wikipedia et FreeBase)	C et R
(Lee *et al.*, 2014)	Ontologie de domaine (Bibliographie)	C et R
(Berlanga *et al.*, 2015)	WS (Bioinformatique sur UMLS et WikiNet)	C
Mimir (Tablan *et al.*, 2015)	WS (Innondations sur DBpedia et Geonames) Ontologie de domaine (médical)	C et R C et R
(Wang & Stewart, 2015)	Ontologie de domaine (Catastrophes naturelles)	C et R

TABLE 1 – Prototypes de RIS

3 Annotation de concepts et de relations sémantiques

Cette partie présente notre démarche pour l'annotation de concepts et de relations sémantiques dans un texte. La méthode présentée a pour origine les travaux menés dans le cadre des projets ANR DYNAMO (Dudognon *et al.*, 2010) et MOANO (Bessagnet *et al.*, 2013; Buscaldi *et al.*, 2013). Nous étendons ces travaux de manière à prendre en compte l'annotation automatique des relations quelle que soit l'ontologie.

Nous définissons une **ontologie** O par l'ensemble C des concepts et par l'ensemble R des relations entre concepts : $O = (C, R)$. On note $R = \{r_\nu\}$ avec $r_\nu = (\nu, \delta, \rho)$ où la relation r_ν de nom ν a pour domaine de classe δ et pour co-domaine de classe ρ. Dans l'ontologie, à chaque concept est associée une liste de termes qui *dénotent* le concept ; on note T l'ensemble des termes pouvant dénoter un concept (c'est-à-dire, des termes dont la présence dans un texte implique automatiquement la présence du concept dénoté). Rappelons que la rédaction de la liste associée à chaque concept de l'ontologie est du ressort d'un spécialiste du domaine. Par exemple, dans une ontologie pour le domaine Topo-carto de l'IGN, on trouve les concepts *commune*, *bâtiment remarquable* et *château*. En plus des relations hiérarchiques classiques (*is-a*) permettant de modéliser que le concept *château* est un sous-concept de *bâtiment remarquable*, il est possible de modéliser des relations sémantiques comme la relation *embellir* entre les concepts *château* et *commune*.

Le processus d'annotation que nous proposons s'appuie sur quatre prédicats.

Pour les **concepts** :

$subsumes(c_i, c_j)$ où $c_i \in C$ et $c_j \in C$, indique que c_i est un ascendant de c_j ou bien $c_i = c_j$;
$has\_label\_c(c, t)$ où $c \in C$ et $t \in T$, indique que c a pour label le terme t.

Pour les **relations sémantiques** :

$has\_label\_r(r, t)$ où $r \in R$ et $t \in T$, indique que r a pour label le terme t ;
$relation(t, c_\delta, c_\rho)$ où $t \in T, c_\delta \in C$ et $c_\rho \in C$, indique la relation révélée par t entre c_δ et c_ρ.

Le **corpus** est un ensemble D de documents. Chaque document d est composé de plusieurs champs f_i. L'ensemble des n champs d'un document $d \in D$ est noté $F_d = \{f_0, \ldots, f_n\}$. L'unité de traitement du texte est le champ : une partie de phrase, une phrase, un paragraphe, voire un

document (par défaut, la portée du champ correspond à la phrase). Plus le champ est large plus le risque d'ambiguïté augmente.

Ainsi, un champ f contient un concept c si et seulement si un terme t dénotant un concept c' existe et si le concept c' est un descendant de c ou si $c = c'$ (c'est à dire, $has\_label\_c(c', t)$ et $subsumes(c, c')$). De plus, un champ f contient une relation r si trois termes du champ dénotent, l'un la relation et les deux autres les concepts du domaine c_δ et du co-domaine c_ρ correspondants à cette relation.

L'annotation automatique de concepts et de relations sémantiques dans un champ f se déroule en deux temps : une première phase procède à l'identification des concepts et des relations potentielles puis, une seconde phase valide ou non ces relations (voir figure 1).

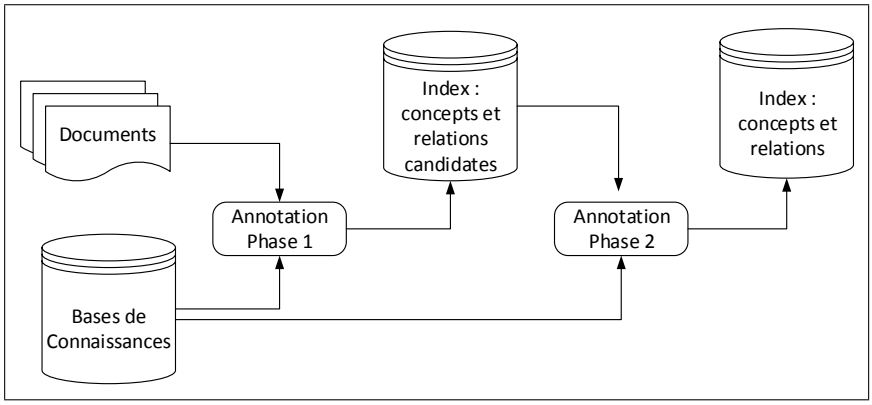


FIGURE 1 – Principe général d'annotation

3.1 Phase 1 : Identification de concepts et de relations

La première phase construit, pour le champ f du document analysé et une ontologie O, les trois ensembles suivants :

- T_f l'ensemble des termes présents dans le champ f du document analysé,
- $AC_f = \{c_0, \ldots, c_m\}$ l'annotation d'un champ f avec les concepts $c_0, \ldots, c_m$ de l'ontologie O,
- $ARP_f = \{r_0, \ldots, r_n\}$ l'annotation d'un champ f avec les relations $r_0, \ldots, r_n$ de l'ontologie O, qui sont potentiellement pertinentes pour le champ f.

Le processus de la phase 1 est décrit par l'algorithme 1 qui, pour chaque terme d'un champ, recherche les éventuels concepts à partir des labels correspondants dans l'ontologie. De la même manière, l'itération suivante de l'algorithme recherche les relations candidates en comparant chaque terme aux labels des relations décrites dans l'ontologie. Il y a ambiguïté lorsqu'un même terme dénote plusieurs relations. L'analyse du domaine et du codomaine d'une relation, dans une seconde phase, permet la différenciation et la validation des relations candidates.

Au terme de cette phase, chaque terme annoté correspond à un ou plusieurs concepts, une ou plusieurs relations, ou encore, un ou plusieurs concepts et relations.

Données :
f : champ étudié,
O : ontologie
Résultat :
T_f : ensemble de termes annotés,
AC_f : ensemble de concepts trouvés,
ARP_f : ensemble de relations sémantiques candidates
début
$T_f \longleftarrow \{\}$
$AC_f \longleftarrow \{\}$
$ARP_f \longleftarrow \{\}$
pour chaque *terme* $t \in f$ **faire**
pour chaque *concept* $c \in C$ **faire**
si $has\_label\_c(c, t)$ **alors**
si $t \notin T_f$ **alors**
$T_f \longleftarrow T_f \cup \{t\}$
$AC_f \longleftarrow AC_f \cup \{c\}$
pour chaque *relation* $r = (\nu, c_d, c_{cd}) \in R$ **faire**
si $has\_label\_r(r, t)$ **alors**
si $t \notin T_f$ **alors**
$T_f \longleftarrow T_f \cup \{t\}$
$ARP_f \longleftarrow ARP_f \cup \{(\nu, c_d, c_{cd})\}$

Algorithme 1 : Phase 1 d'identification de concepts et de relations

3.2 Phase 2 : Validation des relations candidates

La seconde phase a pour objectif de construire l'ensemble de relations sémantiques AR_f à partir du processus de validation appliqué aux relations potentielles contenues dans ARP_f :

– $AR_f = \{r_0, \ldots, r_p\}$ où $r_0, \ldots, r_p$ appartiennent à ARP_f et sont validées pour le champ f.

Ce processus est détaillé par l'algorithme 2. Pour chaque relation détectée lors de la phase 1, il s'agit désormais de valider et d'annoter les triplets « domaine/relation/codomaine ». Ainsi, pour chaque relation candidate, on consulte l'ontologie pour récupérer le concept de *domaine* et le concept de *codomaine* correspondants. Ensuite, on recherche dans le champ, chacun de ces deux concepts parmi les concepts annotés ou leur ascendance.

L'algorithme 2 d'annotation de triplets, qui s'appuie sur la relation comme point de départ, lève la majorité des ambigüités : pour le cas des relations r et r' de mêmes *domaine* et *codomaine*, il validera les deux relations si les vocabulaires associés à r et r' ne sont pas disjoints.

Données :
AC_f : ensemble de concepts trouvés,
ARP_f : ensemble de relations sémantiques candidates
O : ontologie
Résultat :
AR_f : ensemble de relations sémantiques validées
début
  $AR_f \longleftarrow \{\}$
  pour chaque *relation* $r = (\nu, c_d, c_{cd}) \in ARP_f$ **faire**
    si $(\exists c_i \in AC_f, subsumes(c_d, c_i)) \wedge (\exists c_j \in AC_f, subsumes(c_{cd}, c_j))$ **alors**
      $AR_f \longleftarrow AR_f \cup \{(\nu, c_d, c_{cd})\}$

Algorithme 2 : Phase 2 de validation des relations candidates

4 Expérimentation

Nous avons mis en œuvre ces propositions dans une chaîne de traitements dont nous présentons les caractéristiques et des exemples d'expérimentation ci-après.

4.1 La chaîne de traitement

La chaîne de traitement, illustrée sur la figure 2, a été mise en œuvre sur la plateforme GATE (Cunningham *et al.*, 1995; Bontcheva *et al.*, 2004). Elle s'applique à des collections de documents textuels vus comme une suite de phrases. Le premier module intitulé « Traitement de la langue » intègre notamment l'analyseur morphosyntaxique Treetagger (Schmid, 1994) et prend en charge la lemmatisation en langue française afin de tenir compte de variations syntaxiques. Le deuxième module « Annotation d'entités nommées », qui n'est pas instancié systématiquement, permet de détecter des entités nommées décrites dans des bases de connaissances. Le troisième module « Annotation de concepts et de relations » correspond à la mise en œuvre de l'algorithme n° 1 (décrit précédemment). Il s'agit de l'annotation des termes correspondants à des labels définis dans l'ontologie. Chaque annotation comporte les détails suivants : le terme

original, le lemme correspondant, le label identifié, le nom du concept ou de la relation, le type de l'objet reconnu (instance, classe ou relation).

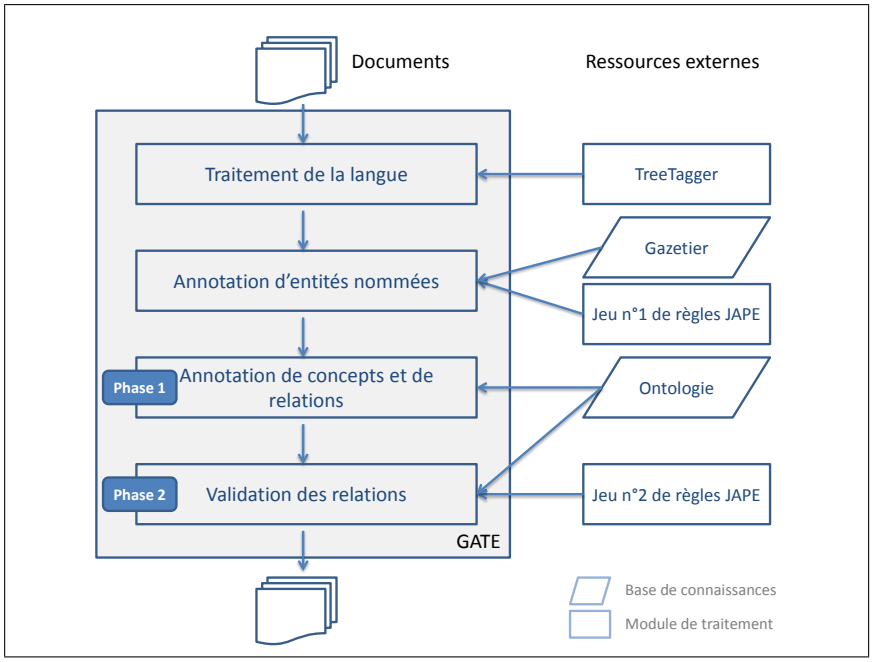


FIGURE 2 – Architecture générale du processus d'annotation sémantique sur la plateforme GATE

Le quatrième module « Validation des relations », quant à lui, correspond à la mise en œuvre de l'algorithme n° 2. Des règles sont définies dans le langage JAPE à partir d'expressions régulières combinant concepts de domaine, de codomaine et relations. Ce module prend, en entrée, les marquages de la phase précédente et s'appuie sur l'ontologie pour valider l'annotation des relations potentielles : à chacune correspond un triplet « domaine/relation/codomaine » défini dans l'ontologie qu'il s'agit de valider à partir des éléments annotés dans le texte. Par exemple, sur la figure 3, la règle CRC0 a pour expression régulière : $C1\ P\ C2$ où $C1$ et $C2$ sont des concepts et P est une relation. $Lookup$ est le résultat d'annotation de la phase précédente comportant notamment un délimiteur de champ. Chaque occurrence de ce triplet $C1\ P\ C2$ est traitée par des instructions Java pour la validation de la relation potentielle. Ainsi, c'est par la présence de concepts relatifs au domaine et au codomaine que la validation permet de lever d'éventuelles ambiguïtés liées aux relations (un même terme pouvant être label de plusieurs relations).

4.2 Trois cas d'étude

Nous avons expérimenté cette chaine de traitement générique avec trois bases de connaissances différentes : les ontologies GEONTO, MOANO et ONTOTHAU, respectivement.

```
Phase: PhaseRelationDetectCRC

Input: Lookup
Options: control =    all

Rule: CRC0
// règle classe relation classe
(
 ({Lookup.type==class}):C1
 ({Lookup.type==property}):P
 ({Lookup.type==class}):C2
 ({Lookup.type==fieldDelimiter})
)
:tripletDPR
-->
:tripletDPR

{
//instructions java calculant et générant les annotations
}
```

FIGURE 3 – Règle Jape pour la validation de relations sémantiques

4.2.1 Ontologie GEONTO

Le projet ANR GEONTO (Mustière *et al.*, 2011) a permis de construire une ontologie de concepts topographiques. Cette ontologie a été conçue par enrichissement d'une première taxonomie de termes, et ce grâce à l'analyse de deux catégories de documents textuels : des spécifications techniques de bases de données de l'IGN et des récits de voyage. Comme décrit dans (Kergosien *et al.*, 2009), l'ontologie GEONTO est une hiérarchie de concepts géographiques et de labels associés. Nous avons enrichi GEONTO avec des relations sémantiques à des fins expérimentales.

4.2.2 Ontologie MOANO

Le projet ANR MOANO (Aussenac-Gilles *et al.*, 2013) a permis de construite une ontologie à partir d'une collection de documents web structurés portant sur le domaine des plantes, non d'un point de vue botanique ou scientifique mais plutôt du point de vue du jardinage. Cette ontologie a été construite de manière automatique par raffinement successifs et contient des relations entre concepts.

4.2.3 Ontologie ONTOTHAU

Le projet CNRS MASTODONS ANIMITEX (Roche *et al.*, 2014) a permis de construite une ontologie, dédiée au bassin de Thau et à l'aménagement du territoire, à partir d'un vocabulaire défini par des experts géographes. Comme pour GEONTO, cette ontologie a été enrichie avec des relations sémantiques à des fins expérimentales.

4.3 Focus sur le cas d'étude GEONTO

Nous illustrons ici la mise en œuvre de notre approche sur le cas d'étude GEONTO.

4.3.1 Exemple de relations sémantiques

La table 2 illustre les exemples des relations Crue, Équipement, Patrimoine et Proximité dont les labels sont des verbes à l'infinitif et les domaines et codomaines sont des concepts.

Relation	Labels	Domaine	Codomaine
Crue	inonder, envahir, recouvrir, emporter...	Cours d'eau	Entité à vocation résidentielle
Équipement	situer, exister, être disponible...	Équipement de loisir	Commune
Patrimoine	situer, exister, embellir, enrichir, mettre en valeur, appartient, ériger, bâtir, construire, comporter, compter...	Élément du patrimoine	Commune
Proximité	croiser, longer, traverser, passer sous, passer sur, surplomber...	Route	Route

TABLE 2 – Exemple de relations sémantiques

Notons ici que les relations *Équipement* et *Patrimoine* pourront être vecteur d'ambiguïté puisqu'elles sont décrites par des ensembles de labels non disjoints.

4.3.2 Exemple d'annotations

La figure 4 illustre un exemple de marquage de la relation sémantique *Patrimoine*, définie dans l'ontologie GEONTO, sur un texte tiré de Wikipedia [3].

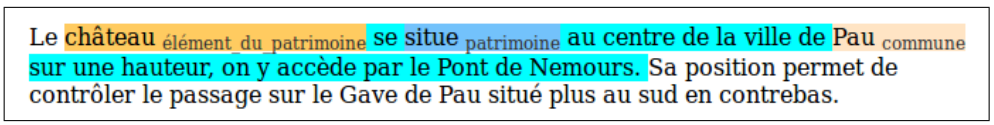


FIGURE 4 – Exemple d'annotation de la relation Patrimoine

L'analyse de ce texte permet d'instancier « château », « situer », « Pau », « Pont de Nemours », « passage », « Gave de Pau »... Or, le verbe « situer » dénote deux relations *Équipement* et *Patrimoine* définies dans l'ontologie. L'ambiguïté est levée par la présence de concepts relatifs au domaine (« château » instancie le concept *Élément du patrimoine*) et au codomaine (« Pau » instancie le concept *Commune*) de la relation *Patrimoine*.

4.3.3 Exemple d'exploitation en RI

Ce marquage nous permet ensuite d'envisager différentes stratégies de RI. Imaginons que nous recherchions tous les documents évoquant des châteaux de la commune de Pau. À cette fin, nous utilisons l'environnement GATE pour mettre en œuvre deux stratégies.

3. http ://fr.wikipedia.org/wiki/Château_de_Pau

L'exemple de la figure 5 illustre une RI basée sur des concepts. Dans ce cas, nous cherchons les concepts « Élément du patrimoine » et « commune ». Deux documents mentionnant « châteaux » et « Pau » sont retournés via la requête. Toutefois, celui mentionnant « le château de Franqueville, visible aisément depuis le boulevard des Pyrénées à Pau » n'est pas pertinent, bien qu'il fasse référence aux concepts « Élément du patrimoine » (château) et « Commune » (Pau).

L'exemple de la figure 6, quant à lui, illustre une RI basée sur des concepts et des relations. Ici, nous affinons la recherche en ciblant la relation Patrimoine. Dans ce cas, un seul document est retourné et il bien pertinent.

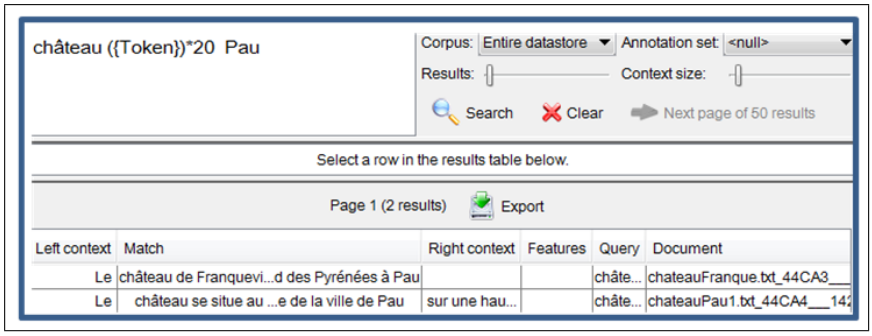


FIGURE 5 – Exemple de RI ciblant des concepts « Élément du patrimoine » et « Commune »

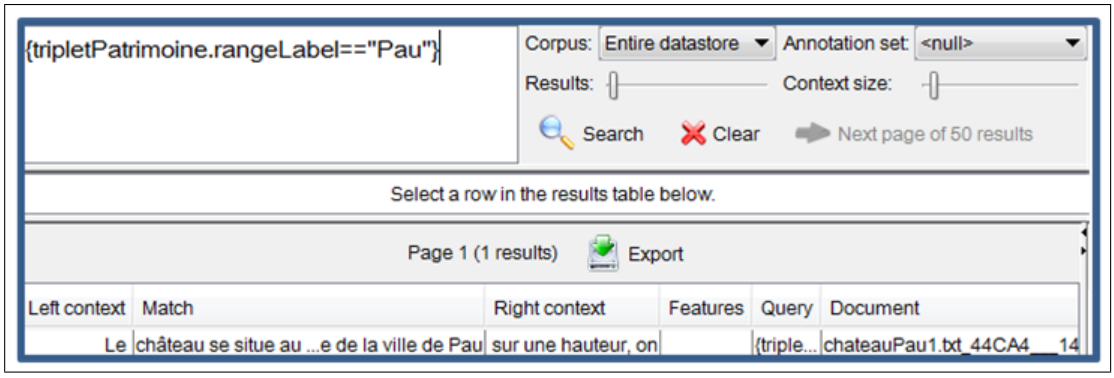


FIGURE 6 – Exemple de RI ciblant des relations « Patrimoine »

5 Conclusion et perspectives

Nous avons décrit une méthode qui vise la reconnaissance automatique, dans des textes, de relations sémantiques (décrites dans une ontologie) entre entités thématiques. Elle s'appuie sur un algorithme de recherche de triplets « domaine/relation/codomaine » qui prend pour point de départ la relation et non les concepts (comme dans la majorité des approches). Notre proposition est générique dans le sens où elle est indépendante du domaine.

Cette proposition est une première étape d'un processus plus large visant la recherche d'information sémantique (RIS). Comme montré précédemment, le processus de RI que nous concevons s'appuie sur les concepts et les relations annotées préalablement. Les travaux de (Maynard & Greenwood, 2012; Buscaldi & Zargayouna, 2013) confirment l'intérêt de telles approches : ils proposent d'indexer les concepts de collections de textes à partir d'ontologies de domaine puis combinent la RI classique de type sac de mots et la RIS de type graphe de concepts. L'approche de RIS que nous proposons s'appuie sur les concepts présents dans les corpus mais aussi sur les relations entre ces concepts (Buscaldi *et al.*, 2013; Bessagnet *et al.*, 2013).

La prochaine étape consistera à évaluer cette approche sur différents corpus de textes et à comparer ces résultats à ceux de nos propositions précédentes ainsi qu'à ceux d'autres systèmes de RIS. Il s'agira de trouver ou de définir un cadre d'évaluation ainsi que des systèmes de RIS ouverts.

Références

ABASOLO J. M. & GOMEZ M. (2000). Melisa. an ontology-based agent for information retrieval in medicine. In *In : Proceedings of the First International Workshop on the Semantic Web (SemWeb2000*, p. 73–82.

AUSSENAC-GILLES N., BUSCALDI D., COMPAROT C. & KAMEL M. (2013). Enrichissement d'ontologies grâce à l'annotation sémantique de pages web. In C. VRAIN, A. PÉNINOU & F. SÈDES, Eds., *Extraction et gestion des connaissances (EGC'2013), Actes, 29 janvier - 01 février 2013, Toulouse, France*, volume RNTI-E-24 of *Revue des Nouvelles Technologies de l'Information*, p. 229–234 : Hermann-Éditions.

BAST H., BÄURLE F., BUCHHOLD B. & HAUSSMANN E. (2014). Semantic full-text search with broccoli. In S. GEVA, A. TROTMAN, P. BRUZA, C. L. A. CLARKE & K. JÄRVELIN, Eds., *The 37th International ACM SIGIR Conference on Research and development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, p. 1265–1266 : ACM.

BERLANGA R., NEBOT V. & PÉREZ M. (2015). Tailored semantic annotation for semantic search. *Web Semantics : Science, Services and Agents on the World Wide Web*, **30**(0), 69 – 81. Semantic Search.

BESSAGNET M.-N., BUSCALDI D., ROYER A. & SALLABERRY C. (2013). Une approche basée sur des relations pour la RI sémantique. In *Atelier Recherche d'Information Sémantique RISE, associé à la conférence IC , édité par Catherine Roussey*, p. 19–33.

BONTCHEVA K., TABLAN V., MAYNARD D. & CUNNINGHAM H. (2004). Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, **10**(3/4), 349–373.

BUSCALDI D., BESSAGNET M.-N., ROYER A. & SALLABERRY C. (2013). Using the semantics of texts for information retrieval : A concept- and domain relation-based approach. In B. CATANIA, T. CERQUITELLI, S. CHIUSANO, G. GUERRINI, M. KÄMPF, A. KEMPER, B. NOVIKOV, T. PALPANAS, J. POKORNÝ & A. VAKALI, Eds., *ADBIS (2)*, volume 241 of *Advances in Intelligent Systems and Computing*, p. 257–266 : Springer.

BUSCALDI D. & ZARGAYOUNA H. (2013). Yasemir : Yet another semantic information retrieval system. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '13, p. 13–16, New York, NY, USA : ACM.

CUNNINGHAM H., GAIZAUSKAS R. & WILKS Y. (1995). *A General Architecture for Text Engineering (GATE) – a new approach to Language Engineering R&D*. Rapport interne CS – 95 – 21, Department of Computer Science, University of Sheffield. `http://xxx.lanl.gov/-abs/cs.CL/9601009`.

DUDOGNON D., HUBERT G., MARCO J., MOTHE J., RALALASON B., THOMAS J., REYMONET A., MAUREL H., MBARKI M., LAUBLET P. & ROUX V. (2010). Dynamic ontology for information retrieval. In *RIAO*, p. 213–215 : CID - Le Centre de Hautes Etudes Internationales D'Informatique Documentaire.

FERNANDEZ M., CANTADOR I., LOPEZ V., VALLET D., CASTELLS P. & MOTTA E. (2011). Semantically enhanced information retrieval : An ontology-based approach. *Web Semantics : Science, Services and Agents on the World Wide Web*, **9**(4), 434 – 452. {JWS} special issue on Semantic Search.

KARA S., ALAN O., SABUNCU O., AKPINAR S., CICEKLI N. K. & ALPASLAN F. N. (2012). An ontology-based retrieval system using semantic indexing. *Inf. Syst.*, **37**(4), 294–305.

KERGOSIEN E., KAMEL M., SALLABERRY C., BESSAGNET M.-N., AUSSENAC-GILLES N. & GAIO M. (2009). Construction et enrichissement automatique d'ontologie à partir de ressources externes. In *Journées Francophones sur les Ontologies (JFO'2009)*.

KIRYAKOV A., POPOV B., TERZIEV I., MANOV D. & OGNYANOFF D. (2004). Semantic annotation, indexing, and retrieval. *J. Web Sem.*, **2**(1), 49–79.

LEE J., MIN J.-K., OH A. & CHUNG C.-W. (2014). Effective ranking and search techniques for web resources considering semantic relationships. *Information Processing and Management*, **50**(1), 132 – 155.

MANNING C. D., RAGHAVAN P. & SCHÜTZE H. (2008). *Introduction to information retrieval*. Cambridge University Press.

MAYNARD D. & GREENWOOD M. A. (2012). Large scale semantic annotation, indexing and search at the national archives. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODIJK & S. PIPERIDIS, Eds., *LREC*, p. 3487–3494 : European Language Resources Association (ELRA).

MUSTIÈRE S., ABADIE N., AUSSENAC-GILLES N., BESSAGNET M.-N., KAMEL M., KERGOSIEN E., REYNAUD C., SAFAR B. & SALLABERRY C. (2011). Analyses linguistiques et techniques d'alignement pour créer et enrichir une ontologie topographique. *Revue Internationale de Géomatique*, **21**(2), 155–179.

NEBHI K. (2012). Ontology-based information extraction for french newspaper articles. In B. GLIMM & A. KRÜGER, Eds., *KI 2012 : Advances in Artificial Intelligence - 35th Annual German Conference on AI, Saarbrücken, Germany, September 24-27, 2012. Proceedings*, volume 7526 of *Lecture Notes in Computer Science*, p. 237–240 : Springer.

ROCHE M., TEISSEIRE M., CRÉMILLEUX B., GANCARSKI P. & SALLABERRY C. (2014). ANIMITEX. analyse d'images fondée sur des informations textuelles. *Ingénierie des Systèmes d'Information*, **19**(3), 163–167.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, p. 44–49.

TABLAN V., BONTCHEVA K., ROBERTS I. & CUNNINGHAM H. (2015). Mimir : An open-source semantic search framework for interactive information seeking and discovery. *Web Semantics : Science, Services and Agents on the World Wide Web*, **30**(0), 52 – 68. Semantic Search.

TRIESCHNIGG R., PEZIK P., LEE V., DE JONG F., KRAAIJ W. & REBHOLZ-SCHUHMANN D. (2009). Mesh up : effective mesh text classification for improved document retrieval. *Bioinformatics*, **25**(11), 1412–1418.

WANG W. & STEWART K. (2015). Spatiotemporal and semantic information extraction from web news reports about natural hazards. *Computers, Environment and Urban Systems*, **50**(0), 30 – 40.

WIMALASURIYA D. C. & DOU D. (2010). Ontology-based information extraction : An introduction and a survey of current approaches. *J. Information Science*, **36**(3), 306–323.

Annotation des Bulletins de Santé du Végétal

Catherine Roussey[1], Stephan Bernard[1]

UR TSCF, Irstea, 9 av. Blaise Pascal CS 20085, 63172 Aubière,France
prénom.nom@irstea.fr

Résumé : Dans cet article nous décrivons les différents schémas d'annotation envisagés pour annoter des bulletins agricoles disponibles sur le web. Notre but est de publier aussi sur le web de données les annotations manuelles permettant le catalogage des bulletins mais aussi les index utilisables par un système de recherche d'information sémantique.

Mots-clés : Annotations sémantiques, annotations spatio-temporelles, recherche d'information sémantique, bulletins agricoles.

1 Introduction

Pour être plus respectueuse de l'environnement, l'agriculture doit modifier ses pratiques, notamment au niveau de l'usage des produits phytosanitaires. Pour ce faire, le plan Ecophyto s'appuie entre autres sur le système de surveillance des pratiques agricoles, dont les Bulletins de Santé du Végétal (BSV) sont un des moyens de communication. Ce corpus du domaine agricole contient des informations sur les attaques des bio-agresseurs des cultures, région par région (par exemple : la DRAAF de la région PACA signale une explosion des attaques de la rouille du blé sur les cultures de blé dur en vallée du Rhône, dans son bulletin du 23 mai 2011).

Nous souhaitons mettre en place plusieurs processus d'annotation spatio-temporelle afin de permettre à des acteurs du domaine agricole de retrouver les BSV répondant à un besoin.

Cet article présente les schémas d'annotation que nous avons définis pour faciliter la recherche des bulletins agricoles sur le web de données. Tout d'abord nous présenterons le corpus des BSV. La section suivante présente un état de l'art des vocabulaires et ontologies que nous avons étudiés pour définir nos schémas d'annotation. Ensuite nous présentons les deux schémas d'annotation que nous proposons pour publier les BSV sur le web de données. Nous décrivons brièvement le premier processus d'annotation manuelle que nous sommes en train de mettre en place.

2 Le corpus des Bulletins de Santé du Végétal

Le Grenelle de l'environnement et le plan Ecophyto ont renforcé les réseaux de surveillance sur les cultures et les pratiques agricoles. Les Bulletins de Santé du Végétal sont une des modalités mises en place par ces réseaux de surveillance. Le Bulletin de Santé du Végétal (BSV) est un document d'information technique et réglementaire, rédigé sous la responsabilité d'un représentant régional du ministère de l'agriculture, tel que la Chambre Régionale d'Agriculture ou encore la Direction Régionale de l'Alimentation, de l'Agriculture et de la Forêt (DRAAF). La figure 1 présente un exemple de BSV de la région Midi-Pyrénées. Ce représentant doit mettre ses bulletins à disposition du public sur son site internet. La conséquence est que les BSV sont

répartis sur différents sites web (un par région). À notre connaissance, il n'existe pas encore de système donnant un accès uniforme à l'ensemble des BSV.

Les BSV sont rédigés en collaboration avec de nombreux partenaires impliqués dans la protection des cultures. La liste des auteurs des BSV varie en fonction de la région et de la filière agricole, ce qui a pour conséquence que leur contenu et leur présentation ne sont pas uniformes et varient en fonction des auteurs. Les BSV diffusent des informations relatives à la situation sanitaire des principales productions végétales de la région et proposent une évaluation des risques encourus pour les cultures. Des données générales concernant les stratégies de lutte (notes nationales, ...) ou sur la réglementation peuvent figurer également dans les BSV. Selon l'actualité sanitaire et/ou la culture, le rythme de parution des BSV est variable, allant d'une parution hebdomadaire à mensuelle. Les BSV sont une synthèse des observations effectuées sur les cultures. Il existe des bases de données d'observations mais la rédaction des BSV oblige leurs auteurs à décider si une observation est un phénomène unique non représentatif ou un phénomène important représentatif d'une réalité. Les BSV ne sont pas une agrégation automatique de données mesurées mais bien une synthèse humaine des jugements sur des observations.

Nous avons récupéré les BSV publiés entre 2009 et 2014 dans 24 régions, soit un peu plus de 15500 bulletins. En moyenne, une région publie plus d'une centaine de BSV par an. Notre but est de constituer une archive pérenne de ces bulletins agricoles afin d'en extraire un ensemble d'information sur les cultures et les niveaux d'attaques de ces cultures au cours du temps. Cette archive sera disponible comme jeux de données sur le web de données. Cette tâche d'archivage fait partie du projet Vespa "Valeur et optimisation des dispositifs d'épidémiosurveillance dans une stratégie durable de protection des cultures", dirigé par l'INRA.

3 Etat de l'art sur les vocabulaires RDF et ontologies utilisés pour l'annotation

Plusieurs vocabulaires RDF et structures de données du web sémantique (ou ontologie) sont proposés pour stocker des schémas d'annotations. Nous présentons dans la section suivante ceux qui ont servi de base à nos schémas d'annotation. L'annotation dans le monde des bibliothèques consiste à associer des données aux documents pour permettre leur catalogage et faciliter leur accès ; on parle alors de métadonnées. L'annotation sur le web consiste à associer à une ressource web une autre ressource (un tag, une note, un autre document).

3.1 DC : Dublin Core

Le Dublin Core est un vocabulaire RDF utilisé dans le monde des bibliothèques pour déclarer les métadonnnées des documents. Il est décrit dans DCMI Usage Board (2012). Ce vocabulaire définit une série de propriétés ("rdf :property") qui, en l'absence de déclarations plus précises, sont interprétées comme des "annotation properties" sur le web de données. La figure 2 présente une partie des propriétés du Dublin Core.

3.2 FOAF : Friend Of A Friend

FOAF, Brickley & Miller (2014), est un vocabulaire RDF définissant les relations (principalement professionnelles) entre personnes. Ce vocabulaire est basé sur un petit ensemble de classes : *Agent*, *Project*, *Organization*, *Document*, *Group*, etc...


BULLETIN DE SANTE DU VEGETAL
MIDI-PYRENEES
Grandes Cultures - n°28
8 juillet 2010
A retenir
MAÏS : Pyrale:
- 1ère génération : vol très étalé avec un pic enregistré entre le 13 juin et le 1er juillet selon les secteurs.
- 2ème génération: pic de vol prévu au cours de la 1ère décade d'août.
Sésamie :
2ème génération : le début du vol devrait intervenir entre le 20 et le 25 juillet, avec un pic prévu entre le 26 juillet et le 03 août.
Maïs
• Stades phénologiques et état des cultures
Les températures de ces derniers jours ont fortement accéléré le développement du maïs. Aujourd'hui, le stade moyen est situé entre 14 feuilles à sortie paniculle, la floraison est proche pour la majorité des parcelles. Les derniers semis arrivent à 10 feuilles.
• Sésamie
Le vol de première génération est terminé. L'observation de pieds de ponte se poursuit. Toutefois ces attaques ne dépassent jamais 1 à 2 % de pieds touchés dans la parcelle. Les piégeages sont peu nombreux pour l'instant.
D'après le modèle, le pic de vol de deuxième génération devrait se situer entre le 26 juillet et le 03 août.
Évaluation du risque :
Le risque faible en premier génération se confirme. Toutefois, si le climat reste sec et chaud, les attaques de deuxième génération pourraient être significatives dans les secteurs les plus touchés en 2009.
• Pyrale
L'étalement du vol de première génération se confirme. Les piégeages restent significatifs sur la majorité des secteurs. Le pic de vol est dépassé dans l'ensemble de la région; du 13 juin au 1er juillet selon les secteurs.
Cependant, les conditions optimales pour les pontes et leur survie se situaient autour du 15 juin, ce qui devrait donner un maximum de vol de la deuxième génération lors de la première décade d'août.
Le début de vol devrait être significatif au 20 juillet dans les secteurs les plus chauds et au 25 juillet pour les secteurs les plus frais.
BULLETIN DE SANTE DU VEGETAL – GRANDES CULTURES N° 28 du 08 JUILLET 2010 – Page 1/2

FIGURE 1 – Un bulletin de santé du végétal de la région Midi-Pyrénées catégorie grande culture

Une personne se définit par un ensemble de "data type properties" : *name*, *age*, etc... Les relations entre personnes sont définies par l'"object property" *knows*, qui peut se spécialiser en fonction des besoins (par exemple, deux personnes créatrices d'un même document sont des *co-authors*). Cette information de création de documents est stockée par le biais de l'"object property" *maker*, entre une personne et le document qu'elle a créé, comme le montre la figure 3. FOAF a aussi été étendu pour stocker des données issues du web social.

3.3 SKOS : Simple Knowledge Organization Schema

SKOS ou Simple Knowledge Organization System (système simple d'organisation des connaissances) est un vocabulaire RDF proposé par le W3C pour représenter les thésaurus, les classifications et d'autres types de vocabulaires contrôlés ou de langages documentaires, W3C (2009).

SKOS permet de stocker les réseaux terminologiques constituant les vocabulaires contrôlés, utilisés entre autres par les documentalistes et les bibliothécaires. La figure 4 est un exemple

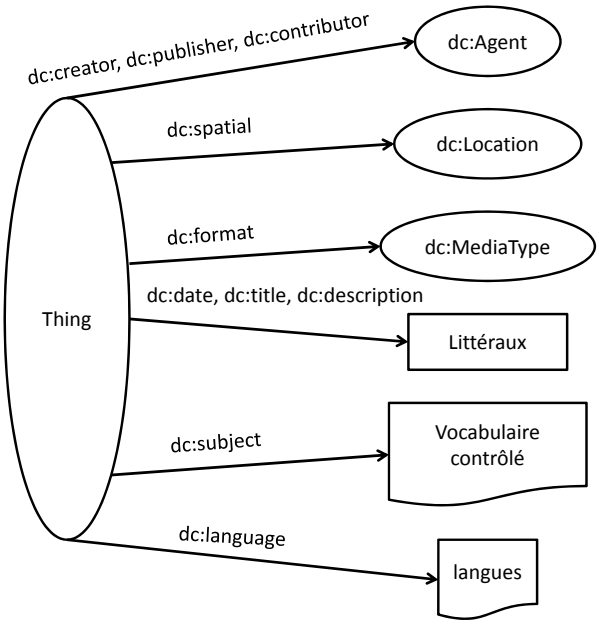


FIGURE 2 – sous ensemble des propriétés définies par le Dublin Core

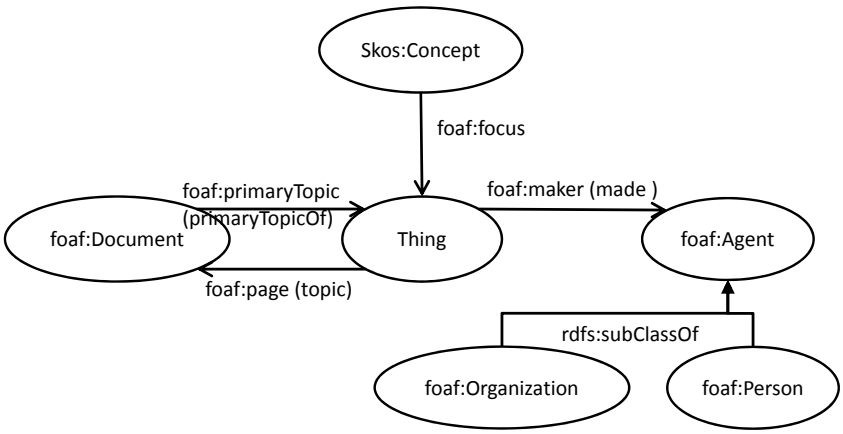


FIGURE 3 – extrait du vocabulaire foaf

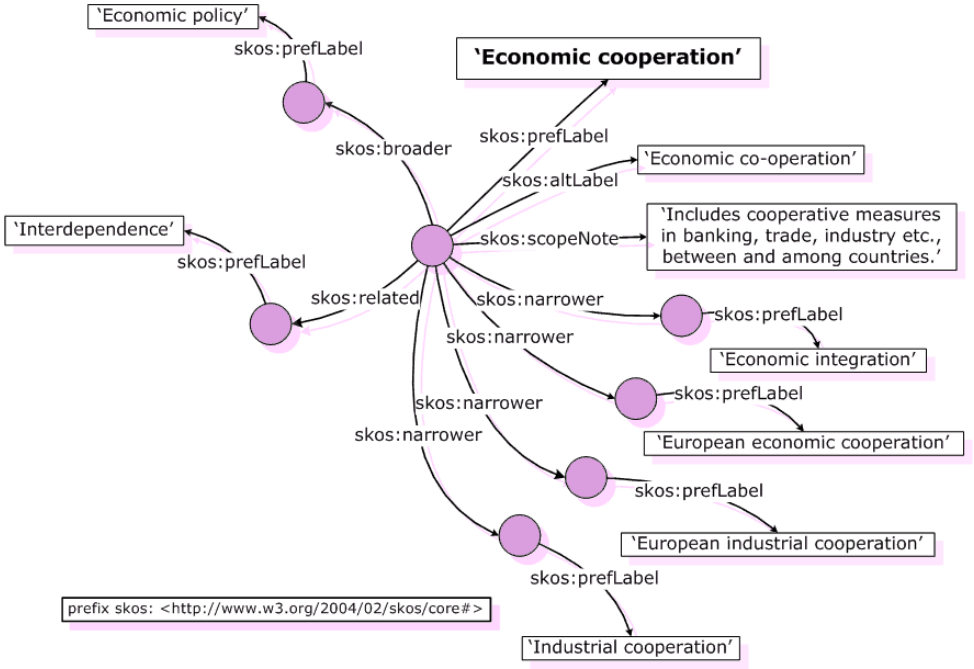


FIGURE 4 – graphe RDF utilisant le vocabulaire SKOS présentant les différents termes liés à "Economic Coopération"

de réseau terminologique issu de W3C (2009). Chaque nœud est un concept SKOS auquel sont rattachés des termes.

3.4 data.bnf.fr

Le schéma d'annotation de la BNF, Bibliothèque Nationale de France (2015), est fondé sur le schéma FRBR (Functional requirements for Bibliographic Records) élaboré par l'IFLA. Comme présenté dans la figure 5, ce schéma comprend trois groupes d'entités liées par des relations :

— les informations sur les documents sont déclarées avec le vocabulaire du Dublin Core,
— les informations sur les personnes physiques ou morales sont déclarées avec le vocabulaire FOAF,
— les informations sur les thèmes sont déclarées avec le vocabulaire SKOS.

Le groupe d'entités qui représente les documents décrit les différents aspects d'une production intellectuelle ou artistique à travers 4 niveaux : l'œuvre, l'expression, la manifestation et l'item.

— Le niveau de l'œuvre est celui de la création intellectuelle ou artistique. Un exemple est l'oeuvre intitulée les Misérables créée par Victor Hugo,
— le niveau de l'expression est caractérisé par la langue, le type de document et les liens de contributions (préfacier, illustrateur, traducteurs...),
— le niveau de la manifestation est celui de la matérialisation d'une expression. Un exemple de manifestation est une édition des Misérables « Nouvelle impression illustrée. 1879-1882. Paris. E. Hugues »,

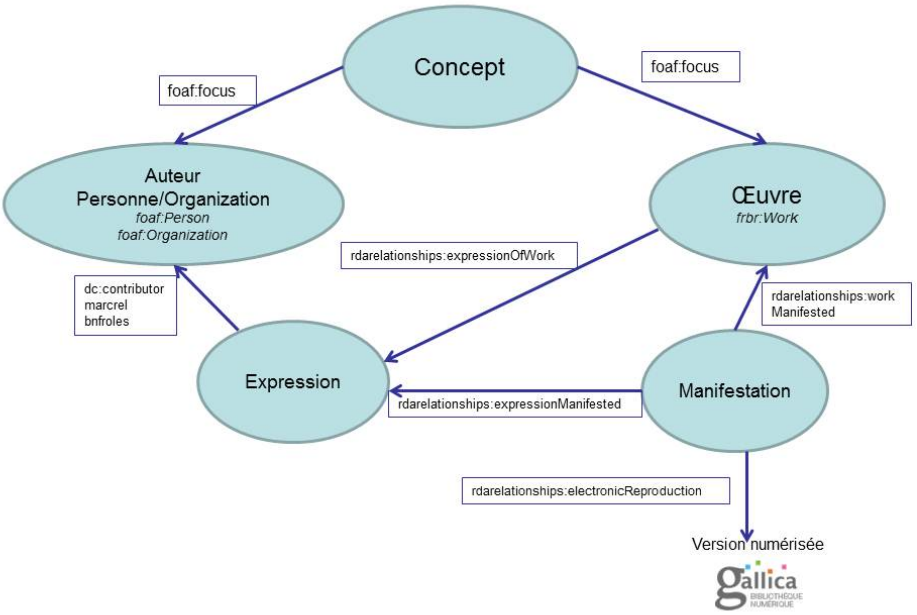


FIGURE 5 – le modèle RDF data.bnf.fr BNF(2015)

— le niveau de l'item est celui de l'exemplaire physique.

Une personne peut être auteur d'une œuvre ou contributeur d'une expression (préfacier, traducteur, librettiste...).

3.5 AO : Annotation Ontology

Cette ontologie est l'un des résultats du projet wf4ever visant à la préservation des résultats expérimentaux. Elle a ensuite donné naissance au projet researchObject pour la publication des ressources scientifiques (article, code, experimentation, etc...) sur le web de données.

Cette ontologie permet d'annoter les documents scientifiques disponibles sur le web à l'aide d'autres ressources, qui peuvent être des mot-clés issus d'un vocabulaire contrôlé (SKOS) ou d'une ontologie du domaine (OWL).

AO permet de préciser si le concept SKOS associé à un mot-clé représente exactement ou approche le contenu de l'annotation, à l'aide des relations *skos :broader* (sens plus générique) ou *narrower* (sens plus spécifique).

Les mot-clés peuvent aussi être une chaîne de caractères proposée par un humain sans contrôle. L'annotation ne se limite pas au «tagage» de document. Dans le contexte du projet wf4ever elle peut aller jusqu'à la prise de note voire la correction collaborative d'un document.

Cette ontologie a été mise en œuvre dans le domaine biomédical et les sciences du vivant (voir Ciccarese *et al.* (2011)). Elle a été utilisée en collaboration avec d'autres ontologies comme PAV qui est une spécialisation de l'ontologie de provenance du W3C pour l'annotation.

3.6 OA : Open Annotation Data Model

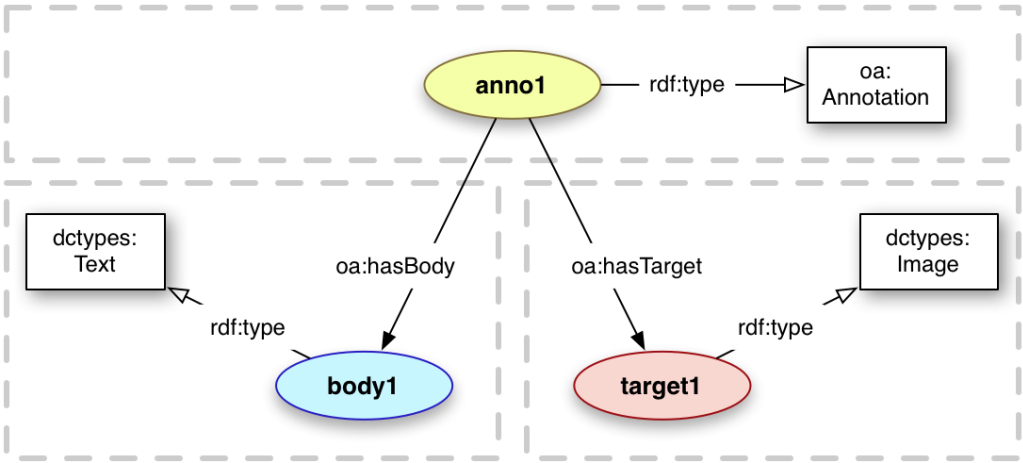


FIGURE 6 – le modèle RDF de base de Open Annotation Core W3C (2013)

Cette série d'ontologies est en cours de développement par un groupe du W3C Haslhofer *et al.* (2014). Les auteurs de AO participent aussi à ce groupe de travail.

L'ontologie Open Annotation Core vise à identifier et décrire les ressources liées à une annotation et à fournir des informations sur la création et l'intention associée à cette annotation ; W3C (2013).

Open Annotation peut être utilisé pour annoter des pages web, éditer collaborativement un document etc... On peut voir OA comme une généralisation et une simplification de AO. Par exemple, OA permet d'exprimer que le contenu de l'annotation est un graphe, sans ajouter plus de détail. Toutefois OA ne donne pas d'indication aussi spécifique que AO sur l'annotation sémantique d'un document web avec une ontologie ou un concept SKOS.

3.7 Synthèse

Nos objectifs sont multiples. Nous voulons tout d'abord proposer un schéma d'annotation permettant le catalogage des BSV afin de faciliter leur recherche. Pour ce faire nous avons choisi de travailler avec des schémas d'annotation standards mis en œuvre par de grandes institutions (BNF).

Nous souhaitons aussi que ces BSV soient utilisés par différents systèmes de Recherche d'Information Sémantique (RIS) et comparer les performances de ces systèmes. Le W3C développe un schéma d'annotation type qui deviendra, s'il est utilisé, un standard. Dans un système de RIS le contenu des documents est représenté par des vecteurs pondérés de concepts, un concept pouvant être soit un concept SKOS issu d'un vocabulaire contrôlé, soit un individu ou la classe d'une ontologie de domaine OWL. Ces vecteurs pondérés sont le résultat d'un processus d'indexation et sont donc appelés index.

Même si AO approche ce besoin, aucune de ces ontologies ne donne de solution pour stocker sur le web de données les vecteurs pondérés de concepts. Nous pouvons noter les travaux de Nešić [Nešić *et al.* (2010)] qui proposent de pondérer les termes utilisés pour l'annotation de document.

Le corpus des BSV sera indexé par différents processus d'indexation issus de plusieurs systèmes de RIS, et nous voulons pouvoir stocker, combiner et comparer ces différents index. Les résultats de plusieurs expériences d'indexation seront disponibles sur le web de données avec le corpus associé. Nous pourrons aussi simuler les résultats d'un système de recherche d'information à l'aide d'un moteur SPARQL en ordonnant les résultats d'une requête.

4 Nos schémas d'annotation

Nous allons proposer deux schémas d'annotation pour les BSVs. Le premier sera un schéma d'annotation pour stocker les métadonnées des BSV comme le ferait un documentaliste, le but étant d'indiquer la date de publication, la région et le type de culture associés à chacun des BSV. Le second sera utilisé pour stocker les index pondérés utilisables par un système de RIS, en étendant l'Open Annotation data model.

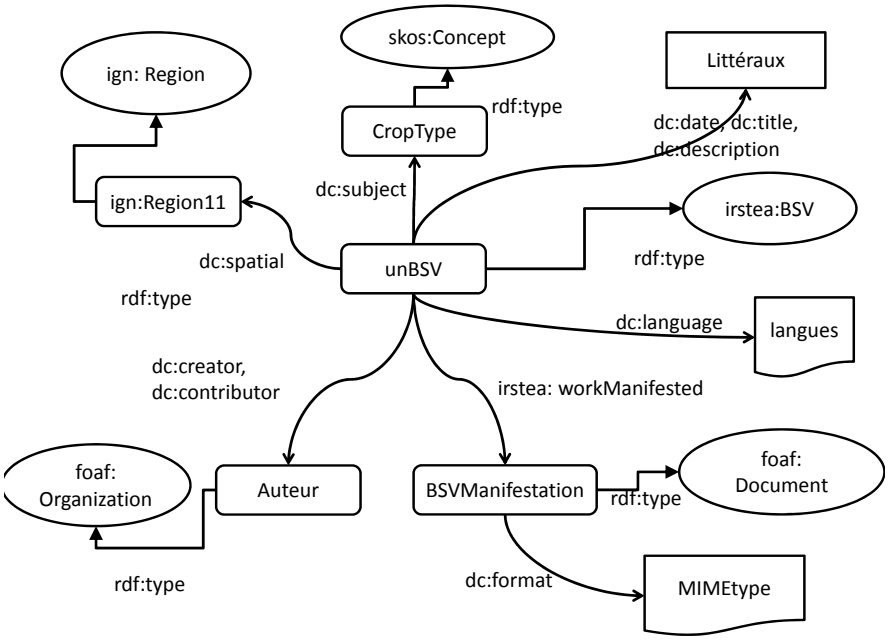


FIGURE 7 – schema d'annotation des BSV

Le premier schéma d'annotation, présenté dans la figure 7 est proche de celui utilisé par la BNF. Nous avons différencié l'entité représentant le BSV comme expression de la création intellectuelle de l'entité représentant sa manifestation physique. Il est en effet possible que différentes copies d'un même BSV soient accessibles sur le web de données avec des formats distincts.

Un bulletin agricole se caractérise par :

1. une métadonnée spatiale correspondant à sa région de publication, indiquée par la propriété *dc :spatial*. Cette propriété lie un bulletin à au moins une région définie dans le jeu de données RDF de l'IGN (`http://data.ign.fr/endpoint.html`).

2. une métadonnée temporelle correspondant à sa date de publication, indiquée par la propriété *dc :date*.
3. une métadonnée thématique correspondant aux types de culture abordées dans le bulletin agricole, indiquée par la propriété *dc :subject*. Cette propriété lie un bulletin à au moins un concept SKOS du thésaurus d'usage des cultures en France que nous avons défini.

L'ensemble de ces données deva être accessible sur le web de données et être utilisable par des moteurs d'inférence. Ce qui signifie que ces données ne doivent pas être enregistrées sous forme d'"annotation properties". Nous devrons définir des "data type properties" et des "object properties" similaires aux "annotation properties" du Dublin Core.

Le second schéma d'annotation a pour objectif de stocker les index produits par différents systèmes de RIS. La figure 8 présente la manière dont OA permet de stocker les informations relatives à la provenance d'une annotation. Ainsi nous pourrons indiquer à quelle date et par qui ont été produits les index, mais aussi quand et par qui ils ont été sauvegardés.

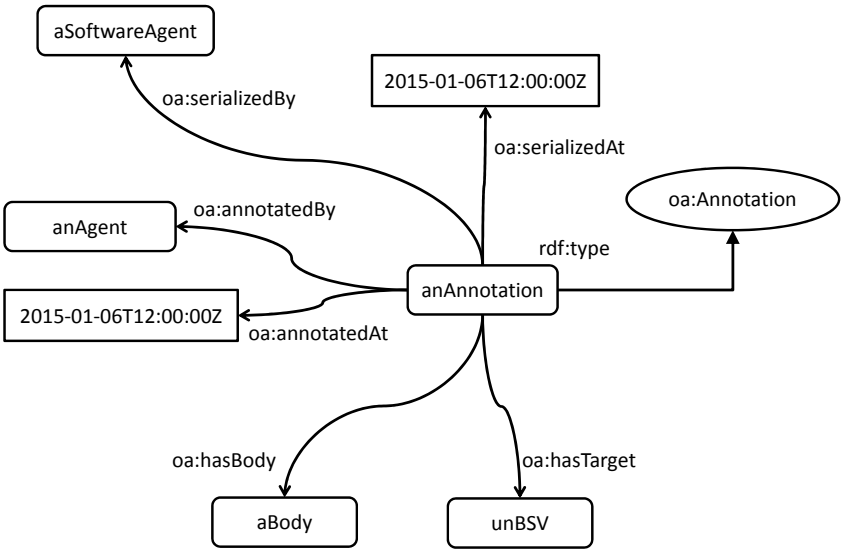


FIGURE 8 – sous-partie de Open Annotation Data Model décrivant la provenance d'une annotation

Nous proposons d'étendre Open Annotation Core pour stocker les index des systèmes de RIS. Cette extension porte le nom de Open Annotation for Indexing (OAI). La figure 9 présente en pointillés les éléments ajoutés à OA et définis par OAI.

Nous définissons d'abord un nouvel objectif d'annotation *oai :indexing*. OA permettant de définir des annotations composites, nous allons définir un nouveau type de tag pondéré représenté par la classe *oai :WeightedTag*. Les tags pondérés sont des éléments d'un individu de type *Composite*. Nous pourrons par exemple utiliser ce schéma d'annotation pour associer non

seulement une région mais aussi les départements de cette région à un BSV. La région et les départements seront les éléments d'une même annotation composite. Le poids affecté à la région et au départements dependra des algorithmes d'indexation.

Concernant les types de culture, nous pourrons, pour chaque type de culture identifié lors du catalogage d'un BSV, associer un sous-ensemble de types de cultures voisines dans le thésaurus des types de cultures. De la même manière que pour les régions, les poids associés aux types de cultures voisines dépendront de l'algorithme d'indexation sémantique.

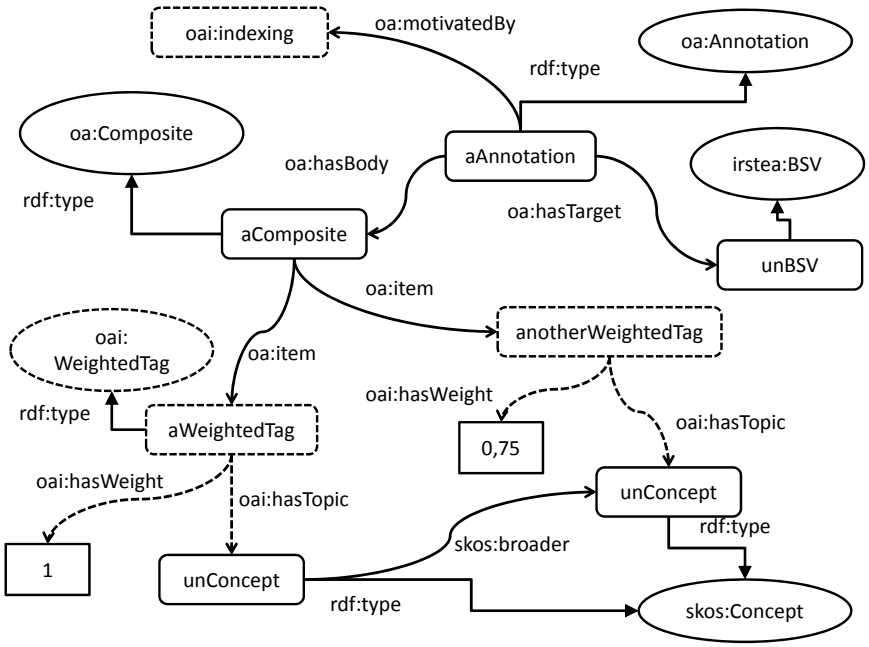


FIGURE 9 – extension de open annotation pour l'indexation

5 Les processus d'annotation

La première méthode d'annotation vise à caractériser tout BSV par au moins sa date de publication, sa région et les types de culture concernés, en utilisant le premier schéma d'annotation.

La description des BSV et leur mise à disposition sur un site web est faite manuellement. Le but est d'extraire semi-automatiquement ces informations, à partir des sites web et des noms de fichiers pdf, pour générer notre premier jeu de données d'annotation. Ces annotations dites "manuelles" sont le résultat de notre travail de moissonnage des BSVs sur le web avec des processus automatiques ou semi-automatiques.

La région est celle de l'administration qui donne accès sur son site web aux BSV. Les sites web que nous avons moissonnés sont en nombre limité. Il nous a donc été possible de récupérer facilement la région associée à un site lors du moissonnage des BSV. Il arrive parfois que les BSV soient le fruit d'une collaboration entre les organismes de deux régions ; deux manifestations distinctes du même BSV existent alors sur les sites web des administrations concernées.

Cette indexation spatiale est faite automatiquement lors de la génération des URI des BSV téléchargés et ne nécessite pas d'intervention humaine.

En ce qui concerne les types de cultures, chaque région publie différentes sortes de bulletins. On trouvera par exemple des BSV sur le colza dans certaines régions, sur les oléagineux dans d'autres, et des BSV sur les grandes cultures dans la plupart des régions de France. Les noms des catégories de BSV ne sont pas normalisés et dépendent des productions principales des régions. En effet, une région peut avoir une catégorie intitulée "petits fruits" alors qu'une autre région l'intitulera "fraises et framboises". L'annotation du type de culture reviendra à associer la catégorie du BSV indiquée sur le site web à au moins une entrée du thésaurus des types de cultures que nous avons défini. Cette indexation thématique est automatisée et se fait à partir d'un patron de transformation construit à la main, qui traduit le nom de catégorie locale en un ensemble de concepts SKOS issus de notre thésaurus.

Obtenir la date de publication n'est pas aussi aisé qu'il n'y parait. Nous avons développé trois processus d'extraction des dates (présentés par ordre de priorité) :

— La date est souvent présente dans le nom du fichier pdf téléchargé. Un premier processus d'extraction à partir des noms de fichiers est réalisé à l'aide de patrons d'extraction de dates typiques.
— La date de création du fichier pdf est aussi présente dans les méta-données du fichier.
— Enfin, nous avons utilisé un processus d'extraction des dates à partir du contenu du fichier pour extraire la date la plus fréquemment rencontrée.

Aucun de ces processus ne permet d'obtenir avec certitude la date de publication du bulletin. Par exemple, certains noms de fichiers ne contiennent pas de date, ou au contraire contiennent une série de chiffres interprétés à tort comme étant une date. Les métadonnées sont parfois illisibles, et il arrive trop fréquemment que le fichier n'ait pas été créé le jour de la publication du BSV (il a été crée la veille, ou corrigé pour être re-créé à une date ultérieure, pas toujours proche). Enfin, le bulletin lui-même contient de nombreuses dates, comme par exemple des dates de relevés ou de mesures, et il est difficile d'identifier avec certitude laquelle correspond à la publication du BSV.

Ces trois processus sont automatiques et nous permettent de sélectionner la date de publication la plus probable selon un algorithme simple : si deux ou trois processus renvoient la même date, c'est celle qui est choisie (73% des cas, soit 11332 BSV sur 15569). Sinon le choix se fera dans l'ordre de priorité décrit ci-dessus (chacun des trois processus pouvant ne retourner aucune date, on sélectionne le premier processus ayant abouti). 0,2% des BSV (c'est-à-dire 37) n'ont pas de date identifiée par ce processus.

L'ordre de priorité a été défini par des statistiques sur les cas où deux dates sur trois sont identiques et par une validation manuelle sur un echantillon de BSV, qu'il conviendra d'étendre pour fiabiliser l'ensemble du processus.

Notre méthode d'annotation dite manuelle effectue une extraction automatique d'informations relatives aux BSV qui ont été publiées par les éditeurs des sites web. Cette méthode va nous permettre de renseigner en partie le schéma d'annotation de catalogage.

Ensuite nous allons pouvoir développer d'autres méthodes d'indexation en utilisant le schéma d'annotation oai. Une méthode serait de transformer et d'enrichir automatiquement les données du schéma de catalogage pour produire des index sémantiques.

Nous espérons par la suite développer une méthode d'indexation capable d'extraire automatiquement des index à partir du contenu des BSV. Cette méthode permettrait de proposer

un second jeux d'index, en particulier pour identifier les agresseurs des cultures et les niveaux de risque. Pour ce faire, nous espérons pouvoir utiliser les sorties du système Vespa Mining Turenne *et al.* (2015).

6 Conclusion

Cet article présente deux schémas d'annotation construits à partir de vocabulaires RDF et d'ontologies. Ces schémas d'annotation ont pour but de faciliter la recherche dans un corpus de bulletins agricoles intitulés Bulletins de Santé du Végétal. Le premier schéma est proche du schéma d'annotation utilisé par la BNF pour le catalogage des documents. Le second schéma basé sur les ontologies Open Annotation Data Model a pour but de stocker les index utilisables par des systèmes de recherche d'information sémantique. Dans des travaux futurs nous devrons valider la mise en œuvre de ces schémas sur une sous-partie du corpus des BSV. Par la suite, nous souhaitons pouvoir combiner et comparer les résultats de différents méthodes d'annotation et d'indexation.

Références

BIBLIOTHÈQUE NATIONALE DE FRANCE (2015). Web sémantique et modèle de données.

BRICKLEY D. & MILLER L. (2014). Foaf vocabulary specification 0.99.

CICCARESE P., OCANA M., GARCIA CASTRO L., DAS S. & CLARK T. (2011). An open annotation ontology for science on web 3.0. *Journal of Biomedical Semantics*, **2**(2).

DCMI USAGE BOARD (2012). DCMI Metadata Terms.

HASLHOFER B., SANDERSON R., SIMON R. & VAN DE SOMPEL H. (2014). Open annotations on multimedia web resources. *Multimedia Tools and Applications*, **70**(2), 847–867.

NEŠIĆ S., CRESTANI F., JAZAYERI M. & GAŠEVIĆ D. (2010). Concept-based semantic annotation, indexing and retrieval of office-like document units. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, p. 134–135 : Centre de hautes études internationales d'Informatique Documentaire (C.I.D).

TURENNE N., ANDRO M., ROSELYNE CORBIÈRE R. & PHAN T. (2015). Open data platform for knowledge access in plant health domain : Vespa mining.

W3C (2009). Skos simple knowledge organization system reference.

W3C (2013). Open annotation data model : Open annotation core.