EGC 2016
REIMS

16ème Conférence Internationale
Francophone sur l'Extraction
et la Gestion des Connaissances

18 - 22 Janvier 2016
http://egc2016.univ-reims.fr

# Actes de l'atelier GAST – **Gestion et Analyse de données Spatiales et Temporelles**

Éric Kergosien (GERiiCO, Université Lille 3)
Thomas Guyet (IRISA-Inria/AGROCAMPUS-OUEST)
Christian Sallaberry (LIUPPA, Université de Pau et des Pays de l'Adour)

http://gt-gast.irisa.fr/gast-2016/

Mardi 19 janvier 2016 Reims

# PRÉFACE

La gestion et l'analyse de données spatiales connaît une dynamique forte grâce au développement de l'estampillage spatial ou temporel des données. Pour répondre aux besoins d'exploration approfondie des données et d'exploitation des informations qu'elles contiennent, des méthodes et outils spécifiques sont requis.

Les objectifs des ateliers GAST (Gestion et Analyse des données Spatiales et Temporelles) concernent notamment la prise en compte de la quantité et de la richesse des données spatiales et/ou temporelles diffusées dans les contenus numériques. La prise en considération de la variété des données numériques (sources, contenus, types de documents, etc.) est également une véritable problématique mais c'est aussi une force dans la quête d'identification de la connaissance. Autrement dit, comment identifier, extraire, structurer et mettre à disposition des acteurs (experts, usagers lambda, etc.) des connaissances s'appuyant sur des données spatiales et temporelles à partir des contenus numériques hétérogènes disponibles ? Ces différents défis lèvent des verrous scientifiques multidisciplinaires qui sont traités au sein la série d'ateliers GAST.

Ces actes regroupent les soumissions acceptées à l'atelier GAST en 2016 dans le cadre de la conférence Extraction et Gestion des connaissances (EGC) organisée à Reims par C. de Runz. Cet atelier est un rendez-vous annuel fédérateur, convivial et scientifique riche de l'ensemble de la communauté s'intéressant à la gestion et à l'analyse de données spatiales et temporelles. Cette année, l'atelier a été organisé en trois temps : une présentation invitée de Danielle Ziebelin sur les données spatio-temporelles ouvertes et liées, puis un ensemble de présentations orales des articles retenus pour l'atelier et finalement un temps dédié à la discussion avec l'ensemble des participants. Nous espérons que le lecteur qui n'a pu y assister trouvera toutes les informations dans les articles de ce volume.

L'article "Identification automatique des types de relations spatiales dans les textes", de Sarah Zenasni, Eric Kergosien, Mathieu Roche et Maguelonne Teisseire apporte un éclairage sur la découverte de connaissances à partir de documents textuels et, plus particulièrement, l'identification d'informations spatiales. La méthode proposée combine des approches de fouille de textes pour identifier les types de relations spatiales de façon automatique. Les résultats des expérimentations réalisées sur un corpus en anglais sont également présentés et discutés.

L'article "Approche pour l'élaboration d'un modèle chronotopique urbain", de Alain Guez et Francis Rousseaux, s'inscrit dans le cadre de l'étude chronotopyque d'un territoire. Ce type d'étude vise à analyser, avec un point de vue géographique, les rythmes de présence et co-présence des résidents – et des habitants temporaires – en fonction des activités, des horaires et l'organisation de la ville. La démarche entreprise dans le cadre de ce travail vise à mobiliser de l'information disponible (ici des horaires d'ouvertures/fermetures d'activité collectée à partir de différences sources)

pour répondre à ces questions de géographie. Il partage ainsi des questionnements sur les représentations conjointes des dimensions spatiales et temporelles. Des premières propositions de représentations de cartes et de caractéristiques temporelles sont faites, mais l'article ouvre surtout sur les défis que représentent leurs questions en terme d'exploitation de données (massives) et de leur analyse automatique.

L'article "Cognisearch Business : un service de recherche d'information d'entreprises sur le web", d'Armel Fotsoh, Annig Le Parc-Lacayrelle et Tanguy Moal, vise la construction de "cartes d'identités" d'entreprises à partir de données du Web. Le modèle d'entité entreprise décrit des informations d'immatriculation (SIREN), des coordonnées (Web, téléphone, adresse) et un contexte (métier, activité, produit). La chaine de traitement comprend des modules de filtrage de sites Web, d'annotation des informations spatiales et thématiques, d'enrichissement des ressources externes, d'indexation et recherche d'information. Un prototype met en œuvre ces propositions.

L'article "La confiance est dans l'air ! Application à l'identification des parcours hospitaliers", de Yves Mercadier, Jessica Pinaire, Jérôme Azé, Sandra Bringay et Maguelonne Teisseire, vise l'identification de séquences fréquentes d'événements ordonnés. Les verrous concernent notamment l'extraction de motifs séquentiels et la notion de confiance appliquée à ces motifs. Le domaine d'application est la prédiction de la trajectoire de patients ayant eu un infarctus du myocarde entre 2009 et 2013. Les résultats obtenus sont discutés par un spécialiste.

L'article "Analyse multi-échelles de référentiels vectoriels via SOLAP", de Marie-Dominique Van Damme et Sébastien Mustière, concerne l'exploration de larges volumes de données hétérogènes. L'approche expérimentale propose d'intégrer une base de données topographique dans un SOLAP. Elle a pour objectif d'analyser et agréger, via des requêtes ad-hoc, les données à grande échelle et vectorielles de l'IGN. Ce travail permet déjà à travers des zooms sur les dimensions spatiales et des sélections temporelles de découvrir différents phénomènes. Ces premiers résultats, via les outils SOLAP, contribuent à la modélisation multidimensionnelle et à l'étude de données vecteurs de l'IGN à des échelles variées.

Ces articles montrent une large étendue des recherches actuelles dans le domaine de la gestion et de l'analyse de l'information spatiale et temporelle. Nous y trouvons avec plaisir des thématiques aussi différentes que le traitement automatique des langues et la fouille portant sur des données temporelles, spatiales et thématiques. Tout ceci correspond aux intérêts premiers de GAST. Au travers de cet atelier, nous espérons que les orateurs, les auditeurs et les lecteurs constatent la complexité que continue de poser l'information temporelle et spatiale, qu'ils voient les défis qui se posent encore aux chercheurs.

Nous tenons à remercier tous les auteurs pour leurs propositions d'articles ainsi que les membres du comité de lecture qui ont su respecter les contraintes imposées par le

planning serré d'un atelier et dont les relectures ont été de qualité pour l'ensemble des articles. Nous remercions également chaleureusement Danielle Ziebelin, professeure à l'Université de Joseph Fourier, pour son intervention en tant que conférencière invitée à la journée GAST'2016. En espérant que ces articles vous apporteront de nouvelles perspectives sur la gestion et l'analyse de données spatiales et temporelles, nous vous souhaitons une bonne lecture.

Eric Kergosien
Université Lille-3/GERiiCO

Thomas Guyet
Agrocampus-Ouest/IRISA-Inria

Christian Sallaberry
Université de Pau/LIUPPA

**Membres du comité de lecture**

Peggy Cellier - IRISA, Rennes
Christophe Claramunt - Ecole Navale, Brest
Géraldine Del Mondo - INSA, Rouen
Thomas Devogele - LI, Tours
Catherine Domingues - IGN, Saint-Mandé
Frédéric Flouvat - PPME, Nouméa
Thierry Joliveau - CRENAM, EVS, Saint-Etienne
Éric Kergosien - GERIICO, Lille
Florence Le Ber - ENGEES, Strasbourg
Simon Malinowski - IRISA, Rennes
Thomas Guyet - AGROCAMPUS-OUEST/IRISA, Rennes
Simon Malinowski - IRISA, Rennes
Nicolas Meger - LISTIC, Annecy
René Quiniou - Inria, Rennes
Sébastien Mustière - IGN, Saint-Mandé
Mathieu Roche - CIRAD, Montpellier
Fatiha Saïs - LRI, Paris
Nazha Selmaoui-Folcher - PPME, Nouméa
Christian Sallaberry - LIUPPA, Pau
Nazha Selmaoui - PPME, Nouméa
Maguelonne Teisseire - IRSTEA, Montpellier
Karine Zeitouni - PRISM, Versailles

## TABLE DES MATIÈRES

### Présentation invitée

Données spatio-temporelles ouvertes et liées : surveillance des ressources en eau
*Danielle Ziebelin* . . . . . 1

### Articles de l'atelier

Identification automatique des types de relations spatiales dans les textes
*Sarah Zenasni, Eric Kergosien, Mathieu Roche, Maguelonne Teisseire* . . . . . 3

Approche pour l'élaboration d'un modèle chronotopique urbain
*Alain Guez, Francis Rousseaux* . . . . . 9

Cognisearch Business : un service de recherche d'information d'entreprises sur le web
*Armel Fotsoh Tawofaing, Annig Le Parc-Lacayrelle, Tanguy Moal* . . . . . 21

La confiance est dans l'air ! Application à l'identification des parcours hospitaliers
*Yves Mercadier, Jessica Pinaire, Jérôme Azé, Sandra Bringay, Maguelonne Teisseire* 35

Analyse multi-échelles de référentiels vectoriels via SOLAP
*Marie-Dominique Van Damme, Sébastien Mustière* . . . . . 47

**Index des auteurs** **58**

# Données spatio-temporelles ouvertes et liées : surveillance des ressources en eau

Danielle Ziebelin*

* Laboratoire d'Informatique de Grenoble, Université Joseph Fourier

## Résumé

Le but des données liées n'est pas simplement de mettre sur un site web un ensemble de données, mais de construire autour de ces données, un environnement capable de les mettre en relations avec différents contextes, avec d'autres sources d'information et d'autres sources de traitement. Dans le cadre des données spatio-temporelles, un certain nombre de standards émanant de l'OGC (Open Geospatial Consortium) et du W3C émergent et contribuent, avec l'utilisation d'ontologies génériques et métiers, à sémantiser les données et à les enrichir. Ces principes seront présentés au travers d'un exemple sur la publication et la liaison de données décrivant des ressources en eau.

# Identification automatique des types de relations spatiales dans les textes

Sarah Zenasni[*,***] Eric Kergosien [**]
Mathieu Roche [*,***] Maguelonne Teisseire [*,***]

[*]UMR Tetis (IRSTEA, CIRAD, AgroParisTech), France
prénom.nom@teledetection.fr
[**]GERiiCO, Univ. Lille 3, France
prénom.nom@univ-lille3.fr
[***]LIRMM, CNRS, Univ. Montpellier, France
prénom.nom@lirmm.fr

**Résumé.** La découverte de connaissances à partir de documents textuels, en particulier l'identification d'informations spatiales, est une tâche difficile due à la complexité de l'analyse des textes écrits en langage naturel. Dans nos travaux, nous proposons une méthode combinant des approches de fouille de textes pour identifier les types de relations spatiales de façon automatique. Les résultats des expérimentations réalisées sur un corpus en anglais sont présentés et discutés.

## 1 Introduction

L'extraction d'information spatiale prend une importance croissante non seulement sur les entités spatiales (ES), mais aussi sur les relations entre entités spatiales. Ces dernières se sont avérées complexes à saisir, à définir et donc à modéliser. Le travail présenté dans cet article se situe dans un tel contexte, l'objectif est de découvrir le type des relations entre les entités spatiales exprimées dans les textes. Nous nous concentrons plus particulièrement sur trois types de relations spatiales : région (par exemple, "leading up"), direction (par exemple, "going up") et distance (par exemple, "near"). La suite de cet article est organisée de la façon suivante. La section 2 présente une brève introduction des travaux existants en extraction d'information spatiale. Puis, nous décrivons, en section 3, les deux approches proposées et leur combinaison. Nous détaillons, en section 4, le protocole expérimental et les résultats obtenus. Finalement la section 5 présente la conclusion et les perspectives de nos travaux.

## 2 État de l'art

De nombreux travaux s'intéressent à l'identification d'Entités Nommées (EN), et plus particulièrement d'Entités Spatiales à partir de données textuelles (Nadeau et Sekine, 2007). Ces approches s'appuient sur des méthodes linguistiques (par patrons d'extraction par exemple) (Maurel et al., 2011) et / ou sur des méthodes statistiques (Velardi et al., 2001). Ces techniques sont intéressantes pour l'identification d'Entités Spatiales, mais elles ne permettent pas

d'identifier l'information spatiale de manière plus exhaustive. Une meilleure représentation de la connaissance spatiale peut être obtenue en considérant les informations sur les relations spatiales. Globalement, les relations peuvent être identifiées par des calculs de similarité entre des contextes syntaxiques (Grefenstette, 1994), par prédiction à l'aide de réseaux bayésiens (Weissenbacher et Nazarenko, 2007), par des techniques de fouille de textes (Grčar et al., 2009). Cependant, ces approches ne permettent pas toujours d'identifier la sémantique de la relation. Nos travaux s'inscrivent dans ce contexte et visent à reconnaitre de façon automatique le type de la relation spatiale étudiée. Plus précisément, nous proposons une méthode hybride, combinant des informations lexicales et contextuelles et une approche de fouille de textes pour prédire le type de relations spatiales.

# 3 Prédiction du type de relation spatiale

Dans la suite de cet article, nous nous appuierons sur les deux phrases ci-dessous pour lesquelles nous cherchons à prédire la classe des relations spatiales (en gras).
**Phrase 1 :** Stairs are **leading up** to the entrance.
**Phrase 2 :** Four locals are **sitting on** a bench in a canteen kitchen , **leaning on** a red brick wall.

## 3.1 Par comparaison de chaînes de caractères

Parmi les nombreuses mesures de similarité existantes, nous avons choisi deux méthodes ***String Matching (SM)*** (Maedche et Staab, 2002) et ***Lin*** (Lin, 1998) qui sont classiquement utilisées dans la littérature car elles produisent des résultats pertinents (Duchateau et al., 2008).

### 3.1.1 String Matching

$SM$ est une mesure lexicale fondée sur la ***distance de Levenshtein*** (notée $L$) (Navarro, 2001), elle calcule la somme minimale du coût des opérations **suppression, insertion, remplacement** nécessaires pour transformer une chaîne de caractères *Ch1* en *Ch2*. À partir de

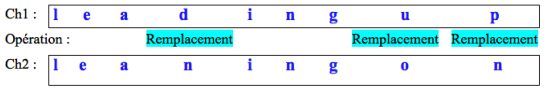



FIG. 1 – *Distance de Levenshtein pour les relations "leading up" et " leaning on".*

l'exemple présenté dans la Figure 1, nous obtenons $L$(leading up, leaning on)=3. En effet, il y a trois opérations permettant de passer de la chaîne "leading up" à "leaning on". Après avoir calculé la distance $L$, nous appliquons la formule (1) pour calculer la valeur de $SM$, normalisée entre 0 et 1.

$$SM(Ch1, Ch2) = max[0; (min(|Ch1|, |Ch2|) - E(Ch1, Ch2))/min(|Ch1|, |Ch2|)] \quad (1)$$

À partir de l'exemple des phrases 1 et 2, $SM$(leading up, leaning on)=$max[0, (10-3)/10] = 0.70$. Sur la base de ces mesures, nous avons retourné pour chaque relation candidate pour

laquelle nous voulons prédire la classe, les similarités obtenues avec l'ensemble des relations de l'ensemble d'apprentissage (voir l'explication détaillée en section 4). Nous déterminons ainsi les $K$ relations les plus proches afin de prédire la classe à associer à la relation candidate (algorithme des $K$ plus proches voisins $KPPV$). Pour les cas particuliers tels que les relations composées de deux mots dont le deuxième mot est une relation spatiale **next to, standing in...**, nous faisons l'hypothèse que ces relations sont du même type que celui des relations **to, in...**

#### 3.1.2 Lin

La mesure $Lin$ est une mesure de similarité fondée sur l'identification des n-grammes de caractères. Généralement, la valeur de $n$ varie entre 2 et 5. En posant $n = 3$ (tri-grammes notée tr), nous obtenons le résultat ci-dessous en reprenant l'exemple de la section 3 :
**tr** (leading up) = {**lea**,ead,adi,din,**ing,ng** ,g u, up} = 8
**tr** (leaning on) = {**lea**,ean,ani,nin,**ing,ng** ,g o, on} = 8
**tr** (leading up) ∩ **tr** (leaning on) = 3
La formule 2 présente la mesure *Lin* normalisée entre 0 et 1 :

$$Lin(Ch1, Ch2) = \frac{1}{[1 + |tr(Ch1)| + |tr(Ch2)| - 2 \times |tr(Ch1) \cap tr(Ch2)|]} \quad (2)$$

À partir de l'exemple des phrases 1 et 2, $Lin$(leading up, leaning on) = $\frac{1}{[(1+8+8)-(2\times3)]} = 0.09$.
Sur la base de cette mesure de similarité, nous avons également appliqué l'algorithme $KPPV$ qui retourne la classe majoritaire pour la relation candidate. Notons que les informations lexicales ne sont pas toujours suffisantes. En effet, deux expressions peuvent être lexicalement éloignées mais sémantiquement très proches. Pour résoudre un tel problème, nous proposons, dans la section suivante, de prendre en compte le contexte des relations pour prédire leur classe.

### 3.2 Par proximité contextuelle

À cette étape, nous faisons l'hypothèse que les mots présents autour des relations (toute la phrase ou les $n$ mots autour de la relation), que nous nommons "*monde lexical*", vont nous permettre d'améliorer l'identification du type des relations spatiales. Nous nous appuyons ensuite sur une approche sac de mots "SDM", nous comparons différents facteurs de pondération : nombre d'occurrences, TF-IDF (Salton et Buckley, 1988) et la confiance (Agrawal et al., 1993) afin de sélectionner celui qui nous permet de construire le monde lexical le plus à même d'identifier le type de relation spatiale pertinent. Sur la base des trois pondérations des mots du monde lexical, nous mesurons la proximité fondée sur le cosinus entre les mondes lexicaux propres aux relations candidates et aux relations de l'ensemble d'apprentissage. Une fois l'ensemble des mesures de proximité calculées, nous appliquons l'algorithme $KPPV$ et nous affectons chaque relation candidate à la classe identifiée comme la plus proche.

### 3.3 Combinaison

Observant que toutes ces approches prises séparément restent imparfaites, nous proposons une méthode combinant les deux méthodes précédentes (par comparaison de chaînes de caractères et par proximité contextuelle). Dans le cadre de nos expérimentations (voir section

4), nous obtenons la liste des relations prédites pour chaque approche. En analysant qualitativement les résultats obtenus, nous remarquons que l'approche par comparaison de chaînes de caractères donne généralement de meilleurs résultats. Cependant l'approche par proximité contextuelle donne des résultats sensiblement meilleurs lorsque les relations se composent de plus de 4 termes. Au regard de cette première analyse, nous faisons l'hypothèse que si les relations se composent de plus de $n$ [1] termes, nous privilégions la proximité contextuelle $Cos$, sinon nous choisissons l'approche par comparaison de chaînes de caractères *SM*. La section 4 décrit les résultats de nos expérimentations menées sur un corpus en langue anglaise.

# 4 Expérimentations

Pour mener à bien ces expérimentations, nous avons choisi un corpus en langue anglaise SPRL (Spatial Role Labeling) (Parisa et al., 2012) qui représente un benchmark reconnu dans le domaine. Le corpus est composé de 1213 phrases annotées. Nous avons procédé à une série d'expérimentations dans lesquelles nous avons fait varier les paramètres susceptibles d'influencer les résultats des mesures de performance : $K$ pour l'algorithme de $KPPV$ et $n$ paramètre de fenêtrage.

## 4.1 Évaluation de la proximité lexicale

Afin d'estimer l'efficacité des différentes méthodes, nous appliquons un processus de validation croisée. Dans notre cas, le corpus est divisé en 3 partitions et chaque partition contient 31 relations (18 régions, 10 directions, 3 distances). Le jeu d'apprentissage est constitué successivement de 2 des 3 partitions et le jeu de test permettant d'obtenir les résultats présentés est constitué de la partition restante. Le tableau 1 représente dans la colonne *String Matching 1* les résultats obtenus en terme d'exactitude (accuracy) à partir de la 1ère série d'expérimentations, en appliquant l'algorithme $SM$ uniquement. La colonne *String Matching 2* représente les résultats obtenus en appliquant la règle présentée en section 3.1.1 de relations composées de deux mots. La mesure de similarité $SM$ donne des résultats satisfaisants comparativement à la mesure Lin quelque soit la valeur de $K$ avec un score de 0.82 d'exactitude.

| K | *String Matching 1* | *String Matching 2* | *Lin* |
|---|---|---|---|
| 1 | 0.77 | **0.82** | 0.75 |
| 3 | 0.74 | 0.79 | 0.73 |
| 5 | 0.73 | 0.76 | 0.69 |

TAB. 1 – *Résultats des mesures $SM$ et $Lin$ en terme d'exactitude.*

## 4.2 Évaluation de la proximité contextuelle

Dans cette série d'expérimentations, la prédiction des relations spatiales est effectuée sur la base de l'approche sac de mots classique avec suppression des "mots vides" [2]. Nous appliquons les deux contextes (toute la phrase, $n$ mots autour de la relation) et nous évaluons l'approche

1. Nos expérimentations ont montré que $n = 4$ donne les résultats les plus pertinents
2. http ://xpo6.com/list-of-english-stop-words/

pour chaque monde lexical avec $K'$ variant de 1 à 5. Dans le tableau 2, nous pouvons constater que le contexte $n$ *mots autour de la relation* donne des résultats supérieurs à ceux du contexte *toute la phrase*. Le monde lexical fondé sur le TF-IDF avec $K' \geq 3$ donne des résultats satisfaisants. Comme conclusion de cette série d'évaluations, le meilleur score (exactitude de 0.67) est obtenu avec $K' = 5$ et $n = 2$.

| K' | | *toute la phrase* | *n termes autour de RS* | | |
|---|---|---|---|---|---|
| | | | n = 1 | n = 2 | n = 3 |
| 1 | nombre d'occurrences | 0.61 | 0.62 | 0.62 | 0.60 |
| | TF-IDF | 0.56 | 0.60 | 0.51 | 0.53 |
| | Confiance | 0.56 | 0.62 | 0.62 | 0.60 |
| 3 | nombre d'occurrences | 0.62 | 0.60 | 0.63 | 0.58 |
| | TF-IDF | 0.45 | 0.63 | 0.66 | 0.63 |
| | Conf | 0.40 | 0.60 | 0.63 | 0.56 |
| 5 | nombre d'occurrences | 0.58 | 0.65 | 0.67 | 0.57 |
| | TF-IDF | 0.40 | 0.64 | **0.67** | 0.66 |
| | Confiance | 0.41 | 0.64 | 0.67 | 0.56 |

TAB. 2 – *Résultats de la méthode de proximité contextuelle (notée $Cos$) en terme d'exactitude.*

### 4.3 Combinaison

Dans cette section, nous présentons les résultats de la combinaison. Nous avons réalisé une série d'expérimentations pour identifier la combinaison de paramètres les plus adaptés, i.e. $K = 1$ pour $SM$ et $K' = 5, n = 2$ pour $Cos$ utilisant le monde lexical fondé sur TF-IDF. Ceci nous permet d'obtenir un score d'exactitude de 0.84. Ainsi, la combinaison des deux méthodes (comparaison de chaînes de caractères et prise en compte du monde lexical) se comporte mieux que chaque méthode individuellement.

## 5 Conclusion

Dans cet article nous avons proposé une analyse comparative de deux approches et de leur combinaison pour l'identification automatique du type des relations spatiales. Nous avons défini un monde lexical pour améliorer la prédiction. Puis nous avons proposé une méthode combinant plusieurs mesures de similarité et pondérations. Nos résultats montrent que la combinaison améliore la qualité de la prédiction. Cela nous permet d'explorer de nouveaux modes d'hybridation afin de tirer le meilleur parti des différentes approches (lexicales et contextuelles). Comme perspective, dans un premier temps, nous voulons étudier la généricité de la méthode. Pour cette raison nous voulons exploiter un corpus contenant 7000 documents de Midi Libre [3]. Dans un deuxième temps, nous voulons étudier l'adaptation de l'approche selon le type de textes traités (presse vs. réseaux sociaux/SMS). Ainsi, nous envisageons d'exploiter un corpus contenant plus de 88.000 SMS [4]. En effet, un tel corpus contient des expressions de

3. http ://www.lirmm.fr/ mroche/ANIMITEX/participants.html
4. http ://88milsms.huma-num.fr

spatialité spécifiques à la communication média et aux réseaux sociaux (par exemple : "je v a montpel"). Finalement, nous souhaitons également nous intéresser à la prédiction de la classe propre aux relations spatiales entre différents types d'entités nommées (personne, organisation, etc).

# Références

Agrawal, R., T. Imielinski, et A. N. Swami (1993). Mining association rules between sets of items in large databases. In *Proc. of Int. Conf. on Manag. of Data (SIGMOD)*, pp. 207–216.

Duchateau, F., Z. Bellahsene, et M. Roche (2008). Improving quality and performance of schema matching in large scale. *Ingénierie des Systèmes d'Information 13*(5), 59–82.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Grčar, M., E. Klien, et B. Novak (2009). Using Term-Matching Algorithms for the Annotation of Geo-services. In *Knowl. Disc. Enh. with Sem. and Soc. Info.*, pp. 127–143.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. of the Fifteenth Int. Conf. on Machine Learning (ICML)*, pp. 296–304.

Maedche, A. et S. Staab (2002). Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Int. Conf. EKAW*, pp. 251–263.

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

Nadeau, D. et S. Sekine (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes 30*(1), 3–26.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Comp. Surv.*, 31–88.

Parisa, K., B. Steven, et M. Marie-Francine (2012). Semeval-2012 task 3: Spatial role labeling. pp. 365–373. ACL.

Salton, G. et C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Inf. Proc. Manage. 24*(5), 513–523.

Velardi, P., P. Fabriani, et M. Missikoff (2001). Using text processing techniques to automatically enrich a domain ontology. In *FOIS*, pp. 270–284.

Weissenbacher, D. et A. Nazarenko (2007). Identifier les pronoms anaphoriques et trouver leurs antécédents: l'intérêt de la classification bayésienne. In *Proc. of TALN*, pp. 145–155.

# Summary

Knowledge discovery from texts, particularly the identification of spatial information is a difficult task due to the complexity of the texts written in natural language. In our work, we propose a method combining two statistical approaches (lexical and contextual analysis) and a text mining approach to identify the types of spatial relationships. Experiments conducted an english corpus are presented.

# Approche pour l'élaboration d'un modèle chronotopique urbain

Alain Guez*, Francis Rousseaux**

*LAA/LAVUE UMR CNRS 7218
118-130 avenue Jean-Jaurès
75019 Paris
guezal@wanadoo.fr
http://www.laa.archi.fr/
**URCA (Université Reims Champagne Ardenne)
UFR SEN, Moulin de la Housse
51100 Reims
francis.rousseaux@univ-reims.fr
http://www.univ-reims.fr

**Résumé.** *Exploration chronotopique d'un territoire parisien* est une recherche en cours soutenue par Paris 2030. Cette recherche se propose d'explorer conjointement deux questions liées à l'expérience et aux dimensions temporelles de la ville contemporaine : comment se forme l'expérience du temps dans une ville ? Comment définir un chronotope et quel rôle peut-il jouer dans l'expérience temporelle de la ville ? L'objectif de cette recherche est de mettre au point, à partir du cas parisien, un modèle descriptif et conceptuel de chronotope urbain à travers une approche croisée des faits objectivables et des expériences vécues. L'élaboration de ce modèle s'appuie notamment sur des bases de données horaires et calendaires des activités de rez-de-chaussée de la ville, qui doivent permettre de déceler des partitions, des rythmes, des périodes différenciées selon les lieux de la ville. L'interprétation de ces bases de données devrait concourir à alimenter une partie du modèle chronotopique recherché.

## 1 Chronotopie et expérience du temps urbain

Nous présentons une recherche en cours soutenue par Paris 2030 et intitulée *Exploration chronotopique d'un territoire parisien*[1].

[1] Recherche financée par la Mairie de Paris dans le cadre de l'appel à projet Paris 2030, et entamé depuis janvier 2015 par une équipe de quatre chercheurs du Laboratoire Architecture Anthropologie (Lavue UMR CNRS 7218) : Alain Guez (porteur de projet responsable scientifique de la recherche), Alessia de Biase (architecte, docteur en anthropologie), directrice du LAA, Federica Gatta (architecte, docteur en urbanisme), chercheuse au LAA , Piero Zanini (architecte, docteur en anthropologie), chercheur au LAA. A cette première équipe originelle ont été associés Stefano Stabilini, chercheur de la Faculté d'architecture du Politecnico di Milano, expert de politiques temporelles urbaines italiennes, et plus récemment Francis Rousseaux, professeur en informatique à l'Université de Reims et chercheur associé à l'Ircam.

Cette recherche se propose d'explorer conjointement deux questions liées à l'expérience et aux dimensions temporelles de la ville contemporaine : comment se forme l'expérience du temps dans une ville comme Paris ? Comment définir un chronotope et quel rôle peut-il jouer dans l'expérience temporelle de la ville ? Ces questions croisent d'emblée des approches différentes et articulées qui nécessitent de mobiliser des compétences et des disciplines variées pour tenter d'y répondre.

Une première matrice descriptive des chronotopes a été définie dès le milieu des années 90 par les chercheurs de la Faculté d'architecture du Politecnico di Milano à l'initiative du professeur Sandra Bonfiglioli (Bonfiglioli, 1997). Cette matrice avait pour ambition d'essayer d'appréhender conjointement les dimensions physiques et sociales des territoires habités à partir d'un ensemble de données temporalisées. Cette approche chronotopique a permis de relever et de révéler des aspects organisationnels qui caractérisent les aires territoriales en fonction des activités localisées dont les horaires et les calendriers participent à construire des rythmes de présence et co-présence des résidents et habitants temporaires. Le sociologue Guido Martinotti (Martinotti, 1993), a introduit cette distinction en identifiant plusieurs populations qui habitent la ville et les territoires selon des régimes temporels différents et qui, ensemble, composent des co-habitations temporaires : les résidents, les « navetteurs », les city-users ou consommateurs métropolitains, les metropolitan businessmen.

L'approche chronotopique développée par le Politecnico di Milano propose également de considérer les lieux comme des constructions historiques, c'est-à-dire d'en reconnaître non seulement les éléments de sédimentation mais aussi de prendre en compte l'histoire des lieux en essayant d'en expliciter les significations actuelles dans une perspective de transmission et d'héritage, d'éléments aujourd'hui signifiants. Une dernière strate d'interprétation, qui n'est pas déliée des précédentes, propose d'inclure dans la définition des aires chronotopiques, les enjeux de mobilité au travers de la reconnaissance des pratiques caractéristiques en terme d'accessibilité et de mobilité des différentes aires territoriales.

Cette matrice descriptive se veut multiscalaire dans la mesure où elle explore à la fois la vie quotidienne à l'aune de l'histoire des territoires et des lieux, mais aussi différentes étendues géographiques. Le quotidien d'un territoire est déployé sur différentes périodes journalières, hebdomadaires, annuelles pour en saisir les rythmes selon les divers cycles et saisonnalités propres aux pratiques qui le caractérisent. La perspective historique épaissit cette compréhension de la ville au présent, à la fois en reconnaissant des permanences morphologiques, fonctionnelles, ou symboliques mais aussi en constatant la plasticité des espaces investis par des pratiques qui peuvent s'actualiser dans des formes héritées et en partie figées dans la « matière urbaine » qu'il s'agisse de voies, d'infrastructures, de parcelles, de bâtiments, ou de fonctions.

L'approche s'est construite entre une réflexion théorique et expérimentale, ancrée dans les politiques temporelles (Bonfiglioli, Mareggi, 1997), enrichie de questionnements provenant de différents champs disciplinaires comme la sociologie, l'urbanisme, la géographie, l'histoire, et aussi la physique, l'astrophysique, les mathématiques, notamment appliquées aux problématiques de mobilité.

On peut rapprocher aujourd'hui ces explorations de propositions précédentes développées notamment par les Time geographers de l'école de Lund, dont Torsten Hägerstrand fut un des pionniers (Hägerstrand, 1970) en introduisant dans la géographie la question de l'analyse des temps individuels spatialisés, par Kevin Lynch (Lynch, 1972) dans sa critique de la ville moderne, ou encore par Henri Lefebvre (Lefebvre, 1992) dans le cadre de son projet ryhtmana-

lytique. Le terme de chronotope n'apparaît pas chez ces auteurs qui tous cherchent pour autant à dépasser la séparation entre espace et temps et à en saisir les articulations que ce soit dans les organisations de la vie individuelle comme chez Hägerstrand, dans la recherche d'expression du temps à travers l'architecture chez Lynch, ou dans la construction d'une posture de rythmanaliste[2] chez Lefebvre.

# 2 Un cas d'étude parisien

Afin d'instruire les questions posées par cette recherche, nous nous appuyons sur un territoire échantillon : une tranche nord-sud de territoire parisien d'1,3 km de large sur 13 km de long située dans le secteur est de Paris. La dimension de cette tranche est liée à l'utilisation d'un maillage de carrés d'1,69 km$^2$ que le Laboratoire Architecture Anthropologie utilise pour ses recherches parisiennes.

Comme le propose Alessia de Biase (De Biase, 2014 106-107), « la dimension de cette grille n'est jamais absolue [ni hors contexte], à la différence de celle, uniforme, projetée sur l'ensemble du globe ». La définition du « pas » parisien s'appuie sur le fait que dans 1,69 km$^2$ se trouvent généralement au moins deux stations de métro ainsi qu'un ensemble de commerces, de services et d'équipements ce qui permet une vie quotidienne locale et un potentiel d'accessibilité métropolitaine.

Les caractéristiques spatio-temporelles de chacun des tissus urbains qui composent cette tranche de territoire sont intimement liées aux fonctions qu'ils accueillent. Celles-ci participent à construire les rythmes d'usage des espaces bâtis et des espaces publics dont les caractéristiques peuvent varier à plusieurs reprises dans une même journée, entre le jour et la nuit, la semaine et le week-end ou au cours des saisons. À ces rythmes réguliers s'ajoutent des moments particuliers comme ceux structurés par des pôles événementiels dont le calendrier construit de fortes intensités momentanées en contraste avec les rythmes réguliers de leur environnement.

Afin de saisir les rythmes de l'espace public parisien nous avons eu la nécessité de disposer de données horaires et calendaires d'ouverture des espaces riverains du domaine public. Nous faisons ici l'hypothèse d'une forte synchronie – dans une ville dense, intense, et avec une forte pression foncière – entre les rythmes de vie des habitants résidents et temporaires, et les horaires et calendriers des activités localisées. Celles-ci peuvent en effet fonctionner comme des attracteurs construisant des polarités de taille et de poids différents selon les cas, et aussi comme des ressources agrégées entre-elles, non seulement par une proximité physique mais aussi par des co-rythmies opportunistes caractérisant différemment les aires urbaines. Il nous est donc apparu qu'à travers les données horaires et calendaires des activités localisées on peut explorer plus précisément ces hypothèses.

Dans un premier temps de la recherche il s'est agi d'identifier des sources d'informations multiples. Nous travaillons actuellement sur quatre sources de données qui nous ont été mises à disposition et qui couvrent des périmètres différents (voir figure 1 ci-après) et de relevés élaborés par notre équipe :

[2] « Le rythmanaliste fait appel à tous ses sens. Il se sert, comme repères, de sa respiration, de la circulation de son sang et des battements de son cœur, du débit de sa parole. […] Sans omettre, bien entendu, le spatial et les lieux, il se rend plus sensible aux temps qu'aux espaces ». Lefebvre, H. (1992), *Éléments de rythmanalyse,* Paris : Syllepse, pp.33-35.

- Des relevés horaires et calendaires élaborées par les membres de notre équipe et couvrant une partie de la tranche du territoire étudié – 1,3 km de large sur 3 km de long – et correspondant à environ 3 000 activités[3] ;
- Les horaires et calendriers relevés par notre équipe ont été associés à la base géolocalisée des activités organisée par l'Atelier Parisien d'Urbanisme (APUR) qui a mis à notre disposition une extraction de ses données – sur le périmètre parisien de notre tranche de territoire – dans le cadre d'une convention ;
- Des horaires d'ouverture journalières sur la semaine, correspondant à un ensemble d'activités de rez-de-chaussée, couvrant toute l'étendue de la tranche de territoire étudié et comprenant environ 2 000 activités relevées[4] ;
- Des horaires d'ouverture journalières sur la semaine correspondant à des activités de rez-de-chaussée sur l'ensemble des arrondissements touchant le territoire étudié, ainsi que sur les communes d'Aubervilliers et Ivry-sur-Seine et comprenant environ 100 000 activités relevées dont 20 000 environ comportent des données horaires et calendaires à l'échelle de la semaine[5].

L'ensemble de ces données nous permet d'une part de mener une analyse à partir de traitements que nous pouvons élaborer avec les outils intégrés dans le Système d'Information Géographique dont nous nous sommes dotés[6] ; et d'autre part, de disposer d'informations horaires et calendaires, toutes géoréférencées, permettant de questionner ces données suite au terrain anthropologique que nous sommes en train d'engager. Il s'agira en effet de vérifier les éventuelles corrélations entre l'expérience vécue par nos interlocuteurs[7] et les agencements spatio-temporels de l'environnement urbain dans lequel ils résident ou travaillent et de préciser en même temps les paramètres qui participent de leur expérience du temps dans la ville. Ces paramètres nous permettront de construire une définition qualitative du chronotope urbain parisien, en intégrant les caractéristiques horaires et calendaires identifiées ainsi que d'autres données, en fonction des résultats de l'approche anthropologique.

Par ailleurs, les données fournies par l'Apur, nous ont permis de définir sur la tranche de territoire étudié, à partir du classement fait par l'Apur des activités de rez-de-chaussée en 221 catégories, des cartes des fréquences d'usage (figure 2), de durées de fréquentation, et des modalités d'accessibilités des rez-de-chaussée de la ville.

Ainsi, indépendamment des horaires et des calendriers des activités localisées, nous sommes en mesure de cartographier les activités en fonction de leur fréquence d'usage (quotidien, occasionnel, exceptionnel), de la durée de fréquentation de ces activités accessibles au

[3] Ces relevés ont été effectués par des stagiaires de notre Laboratoire entre avril et juin 2015.

[4] Ces données ont été fournies par une première entreprise qui les a relevés et qui a mis à notre disposition une extraction de sa base.

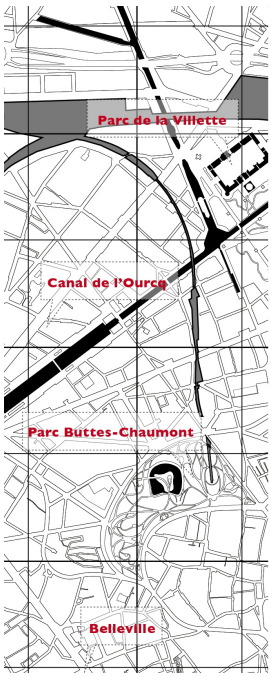
[5] Ces données ont été fournies par une deuxième entreprise qui les a relevés et qui a mis à notre disposition une extraction de sa base.

[6] Le logiciel Arcgis™ a été choisi en accord avec notre partenaire de la faculté d'Architecture du Politecnico di Milano.

[7] Une vingtaine d'entretiens est prévue avec des interlocuteurs de différents âges, rythmes de travail, lieux de résidences ou de travail. Un échantillonnage a été défini afin de recueillir des expériences significatives et variées de la ville.

public (courte, moyenne, longue), de la modalité d'accessibilité de ces activités (libre, sur rendez-vous, ciblée, nécessitant une transaction économique).

Périmètre de l'étude



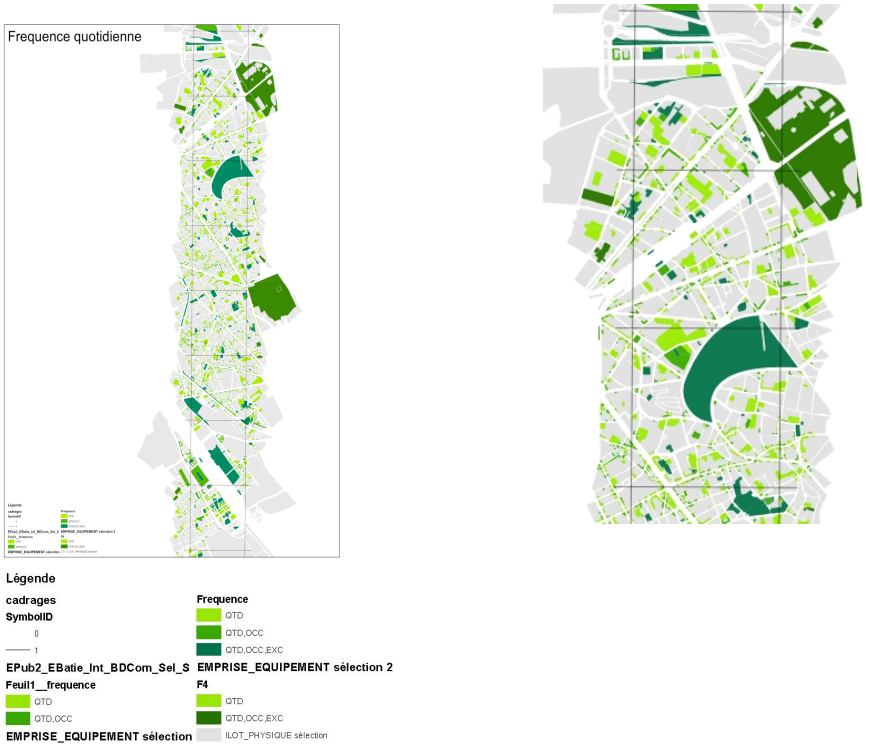

Périmètre de relevé horaire et calendaires réalisé par notre équipe.

Données horaires et calendaires pour environ 2 000 activités géolocalisées

Données horaires et calendaires pour environ 3 000 activités géolocalisées et rapprochées des données Apur

FIG 1 - Périmètre des différentes bases de données géolocalisées et datées des activités de rez-de-chaussée de la ville.

Ces catégories qualitatives sont quantifiables et localisables dans l'espace et nous permettent de qualifier des aires urbaines en fonction des pratiques possibles de l'espace public et de l'extension de ce dernier dans des espaces plus ou moins perméables au public. Le tissu urbain parisien et son cadre bâti offrent une importante surface commerciale et de service au rez-de-chaussée de la ville. Cette surface apparaît comme une extension praticable du domaine public – au sens foncier et juridique du terme – et participe, par la diversité de l'offre, à la composition d'une vaste gamme d'opportunités offertes aux parisiens, résidents et habitants temporaires, en termes de pratiques urbaines.




Ensemble du périmètre parisien Carte des fréquences d'usage Détail

FIG 2 - Carte des différentes fréquences d'usage des types d'activités localisées au rez-de-chaussée de la ville. Élaboré à partir des données fournies par l'Atelier Parisien d'Urbanisme (APUR). Abréviations de la légende : QTD = quotidien, OCC = occasionnel, EXC = exceptionnel.

# 3 L'identification de configurations temporaires

L'analyse des données horaires et calendaires fait apparaître une variation d'ouverture des activités localisées selon les heures et selon les jours de la semaine. Une première approche a consisté à rechercher les variations de l'ouverture des activités de rez-de-chaussée de la tranche de territoire étudiée au cours de la journée et sur l'ensemble du périmètre de 13km. La journée de jeudi a été choisie pour approfondir cet aspect dans la mesure où c'est une journée qui présente l'offre diurne la plus importante[8].

[8] Des explorations sont également en cours sur l'offre nocturne évidemment plus importante les vendredi et samedi et apparaissant dans certains secteurs de la tranche étudiée. A l'échelle de l'année, l'offre de fin de semaine va également présenter des variations importantes à certaines périodes par exemple avant les fêtes de fin d'année où on peut observer une plus importante offre commerciale les dimanches.

Au cours de cette seule journée, l'offre de rez-de-chaussée de la ville présente 13 moments majeurs d'ouverture et 14 moments majeurs de fermeture des activités comme cela apparaît clairement sur le graphique ci-après (figure 3). Comme on peut le voir, les principaux moments d'ouvertures et de fermetures des activités ne se superposent pas systématiquement, ce qui implique, en terme de configurations de l'offre, que sur la journée du jeudi, les 24h sont découpées en 24 séquences différentes dessinant une partition temporelle singulière.

Cette analyse porte sur l'ensemble des 13km étudiés et selon les aires urbaines, ces résultats pourront varier soit en présentant de fortes synchronicités entre les activités (comme on peut l'observer par exemple dans un secteur de forte offre commerciale autour du marché d'Aligre, où une grande majorité des commerces est synchronisée sur les horaires et calendriers de cet important marché parisien), soit en présentant une grande mixité d'horaires répondant localement à la forte mixité de modes de vie des résidents et des habitants temporaires.

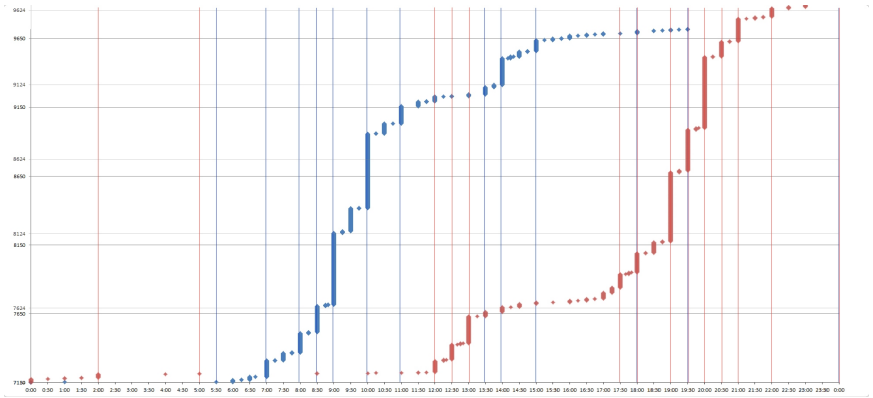


FIG 3 - Graphique représentant les ouvertures et les fermetures des activités de rez-de-chaussée. En abscisse les 24h sont divisées en 48 demi-heures et en ordonnée, les d'activités cumulées s'ouvrant (en bleu) ou se fermant (en rouge) à la même heure.

L'analyse des horaires des activités permet de repérer une partition qui va se singulariser en fonction des lieux de la ville. Lorsqu'on génère par exemple la carte des activités ouvrant un jeudi entre 5h30 et 6h30 du matin on observe que l'avenue de Flandres, dans le XIXème arrondissement, commence à se dessiner comme on peut le voir sur la figure 4. Cette offre matinale se localise principalement le long de l'avenue où se situent les stations de métro donnant accès au territoire métropolitain. Il s'agit principalement de cafés qui se greffent sur les parcours des habitants se déplaçant depuis et vers les stations de métro. Ce genre d'analyse pourra être étendu à l'ensemble de la ville et au-delà, afin de faire apparaître des secteurs du petit matin, du soir, de la nuit ou du dimanche.

Dans certains secteurs caractérisés par une importante présence de bureaux, on pourra par exemple observer une forte offre de restauration à midi qui pourra être exclusivement ouverte pendant cette période en semaine ou qui pourra présenter une amplitude plus importante pour une offre destinée aux habitants résidents. Ceci est vrai par exemple lorsqu'on observe une forte mixité entre résidence et bureaux où les mêmes cafés et restaurants s'adressent alternativement aux travailleurs (le midi en semaine) et aux résidents (le soir et en fin de semaine).

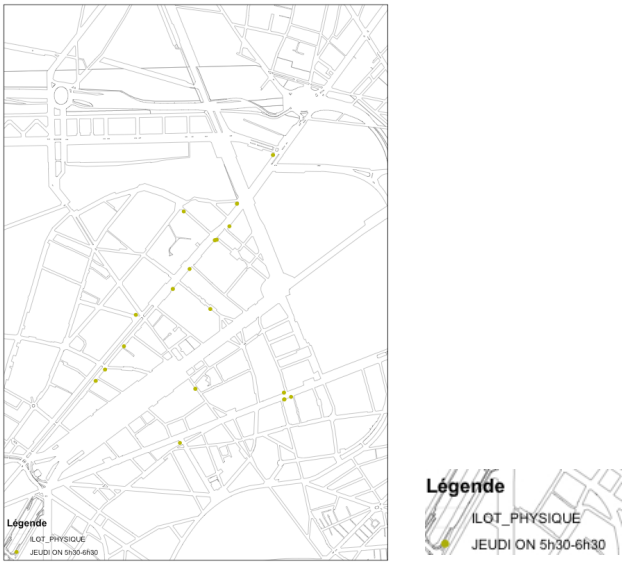



FIG 4 - Carte des activités ouvrant un jeudi entre 5h30 et 6h30 du matin secteur de l'avenue de Flandres, dans le XIXème arrondissement de Paris.

# 4 Une multiplicité d'approches en arrière-plan de l'étude

En avançant dans notre recherche et en croisant nos réflexions avec des chercheurs d'autres domaines, nous nous apercevons que notre recherche nécessite de mobiliser des compétences complémentaires pour exploiter les données dont nous disposons dans la perspective de répondre aux questions qui sont les nôtres. Au-delà de la définition conceptuelle d'un chronotope, qui nous intéresse dans le cadre de la recherche urbaine, comme nous le préciserons ultérieurement, la question de la représentation conjointe des dimensions spatiales et temporelles des phénomènes urbains fait partie intégrante de nos explorations.

Différentes approches de la définition et de la représentation des chronotopes ont pu être menées à ce jour mais n'ont pas encore été codifiées – comme cela fait l'objet de recherche en musique ou en danse notamment avec le travail de Laban (Louppe et alii, 1994) - bien que de nombreux prototypes ont pu être explorés notamment par notre propre Laboratoire. L'analyse des partitions temporelles des lieux habités peut se faire à partir de différents supports allant de la carte aux graphiques, en passant par les chronocartes animées (Guez et Mareggi, 1997 ; Van Schaick, 2011).

Une première famille de cartes, développée à partir des recherches de l'école de Lund, explore la possibilité de représenter des "géographies temporelles" en faisant apparaître les données horaires des parcours quotidien sur une base axonométrique qui utilise l'axe z comme axe du temps (Hägerstrand, 1970).

Une deuxième famille de cartes considère la distance comme un temps nécessaire pour relier deux points du territoire et travaille à montrer comment la proximité géographique se déforme en fonction des temps de parcours (cf. Cauvin, 1994). Ces représentations passent

par des déformations géométriques telle que l'anamorphose ou le dessin de courbes isochrones d'accessibilité des points d'un territoire en fonction des temps d'accès.

La troisième famille, développée dans les recherches des chronotopes du Politecnico di Milano, met l'accent sur l'identification de phénotypes en associant l'analyse des fonctions urbaines, de leurs horaires et calendriers, de leur intensité d'usage et des déplacements des usagers, et des strates historiques de construction du système urbain. Les chronotopes issus de cette approche impliquent une représentation qui se base sur la qualification de différents types de tissus urbains selon la composition de ces paramètres complexes. Un des prototypes de cette approche a été élaboré pour la ville de Pesaro dans le cadre du plan des temps et des horaires de la ville (Bonfiglioli, Zedda, 1999).

La quatrième famille, développée par le Politecnico de Milano et par notre laboratoire, concerne une représentation des espaces en termes de « on/off », c'est-à-dire d'ouverture et clôture au public, et d'intensité des usages sur les 24 heures, sur la semaine et sur l'année. La représentation de ces informations peut être faite à travers une représentation axonométrique similaire à celle de l'école de Lund dans laquelle visualiser les horaires d'ouverture des espaces, ou à travers des « cartes animées » qui montrent une succession d'espaces accessibles à différents horaires/jours/mois (Guez, Stabilini, Zedda, 2000).

Nous nous interrogeons également à ce stade de la recherche sur la meilleure façon de transcrire les caractéristiques temporelles de la ville – rythme, cadence, cycles, variations d'intensités – ce qui nous a naturellement mené à nous rapprocher de l'Ircam dans la perspective d'explorer à travers leurs savoirs et expertises, les partitions et les structures temporelles enfouies dans les données horaires et calendaires dont nous disposons. La traduction en son « sonification » de nos données peut également apparaître comme une nécessité pour percevoir (entendre) les rythmes de la ville que la cartographie « classique » peine à représenter.

# 5 Pistes vers une modélisation

Il y a une forte corrélation entre formes urbaines, formes bâties, distribution fonctionnelle, et offre de services, d'équipements et de commerces situés principalement à Paris, au rez-de-chaussée de la ville. Cette offre, véritable ressource urbaine, ne devient accessible qu'à partir du moment où les agendas et les calendriers individuels et collectifs s'accordent, associant intimement disponibilité spatiale et temporelle.

Notre recherche souhaite vérifier l'hypothèse selon laquelle la qualité de la vie urbaine dépend aussi de la possibilité de bénéficier d'un environnement avec lequel on puisse individuellement et collectivement s'accorder temporellement en pratiquant probablement plusieurs étendues spatiales qui selon les modes de vie et les périodes s'articulent entre-elles.

Le cas parisien nous intéresse particulièrement dans la mesure où il présente une importante complexité due à la variété des ressources régulières et temporaires qu'il offre et à la diversité des modes de vie qu'il accueille et qui dans certains secteurs, notamment ceux situés dans la tranche de territoire que nous étudions, présentent une très forte articulation entre des rythmes différents. La ville apparaît ainsi comme l'imbrication entre des pratiques régulières, occasionnelles, exceptionnelles, simultanées, alternées, synchronisées, ou encore à contretemps les unes par rapport aux autres. Ce sont les conjugaisons, les disjonctions et plus généralement les partitions spatio-temporelles que nous cherchons à repérer et à comprendre, à travers notamment les chronotopes qui singularisent des aires urbaines.

On peut considérer que le chronotope, dont une première matrice descriptive a été rappelée précédemment, peut être le concept générateur d'une modélisation qui tienne compte de l'expérience temporelle des lieux de la ville. En ce sens on pourrait parler d'un modèle spatio-temporel anthropocentré, c'est-à-dire qui s'appuie sur l'expérience humaine pour décrire le monde.

Les données horaires et calendaires dont nous disposons doivent pouvoir être interprétées de sorte à faire apparaître des ensembles de points (activités localisées) qui se comportent temporellement de manière analogue, par exemple qui ouvrent et ferment en même temps à l'échelle du jour et de la semaine. Dans une autre perspective, nous avons pu repérer que le tissu urbain parisien présente une grande diversité d'offre avec, par exemple dans un secteur limité, une variété entre des activités co-rythmiques, mais aussi des activités désynchronisées qui permettent justement une complémentarité, étant ouvertes et disponibles à contre-temps des autres. Nous pouvons faire l'hypothèse, que la vie locale s'appuie justement sur des co-rythmes et des contre-rythmes qui, ensemble, composent un environnement non homogène, mais offrant des complémentarités et des alternatives possibles. Conjointement à ces rythmes réguliers de la vie quotidienne, se concentrent dans certains périmètres des lieux d'intensité, à contretemps des rythmes dominant, constituant des centralités momentanées et ciblées qui participent de la diversité des modes de vie possible qu'offre Paris.

La quantité de données dont nous disposons rend difficile ce repérage s'il n'est pas généré par un moteur de traitement adéquat pour la recherche de *patterns* qui nécessitent de fabriquer *ad hoc* les outils de traitements de nos données.

Nous avons pointé la notion toute deleuzienne d'*agencement spatio-temporel* (Deleuze, et Guattari, 1980), qui évoque aussitôt les *hétérotopies* foucaldiennes (Foucault, 1984) ou les *parcours* de Walter Benjamin (Benjamin, 1989). Reste que les chronotopes qui nous intéressent n'ont de sens que chorégraphiés par les habitants (Berque, 2000).

Et nous aimerions pouvoir conjuguer les contiguïtés spatiales, les rythmes temporels avec le déploiement des activités humaines dans leurs logiques spatiotemporelles, considérées comme faisant un système complexe capable de faire saillir des formes choré-chrono-topiques.

La situation ressemble à celle pointée par Piaget (Piaget, 2012 ; Piaget et Inhelder, 1959), conduit à inventer la notion de « collection figurale » pour faire droit à l'indétermination irréductible qui environne l'enfant, à une période caractéristique de son développement psychique, lorsqu'on lui demande de classer un objet au sein d'un regroupement d'autres objets. L'enfant qui séjourne dans ce stade de développement hésite entre opérer des rapprochements topologiques – typiquement en acceptant de compléter la figure géométrique que préfigure la disposition actuelle de la collection – et opérer des rapprochements aspectuels – en refusant typiquement de regrouper le chat miniature avec les collections de chiens, en dépit du fait que les chiens sont disposés en carré auquel il pourrait être tentant d'ajouter l'angle manquant.

À ce stade de développement – que le très kantien Piaget considère comme pré-catégoriel – l'enfant fait droit de manière indifférenciée et concurrente aux diverses possibilités de regroupement spatial, temporel ou aspectuel. Quant à nous, nous voudrions pouvoir séjourner sans préjugé dans cette indétermination radicale et voir s'agencer ainsi, « équitablement » si l'on peut dire, une collection dynamique faisant émerger des choréchronotopes.

Il s'agirait de généraliser la situation du collectionneur (Wajcman, 1999 ; Rousseaux, 2006) qui – dans l'espace s'il dispose ou accroche des œuvres picturales ou des installations, dans le temps s'il séquence un défilé de mode ou de carnaval – met en scène des tensions,

voire des conflits, entre similarités aspectuelles et contiguïtés spatiotemporelles. Dans notre cas les contiguïtés spatiotemporelles prendraient forme de schéma d'activités articulées, caractérisant des aires urbaines par des *tempos* et compositions singulières.

En dotant les schémas d'activité d'une certaine ouverture possibiliste et les contiguïtés spatiotemporelle d'une certaine élasticité et/ou d'effets de seuil, il serait envisageable de faire saillir des « quartiers émergents », à contour plus dynamique que dans leur acception abstraite ordinaire (« le quartier des affaires », « le quartier des noctambules ») et que dans leur acception topologique locale (« mon quartier »).

# 6 Conclusion provisoire

Si les objectifs de notre recherche sont conceptuels et méthodologiques, ils visent également à construire de nouvelles connaissances et représentations de la ville, proches de l'expérience qualitative de ces habitants.

Dans cette perspective, nous envisageons de soumettre notre analyse des partitions et l'identification des chronotopes urbains à des experts de la ville – urbanistes, programmistes, historiens, … dans la perspective de vérifier avec eux la pertinence de nos catégories et interprétations et d'éventuellement faire évoluer nos modèles conceptuels pour qu'ils adhèrent mieux à l'expérience habitante et qu'ils permettent aussi de doter, les concepteurs et décideurs de la ville de demain, d'outils permettant d'intégrer ces dimensions qualitatives dans la planification et la programmation des espaces de la ville en transformation.

# Références

Bakhtine, M. (1978), Esthétique et théorie du roman, Paris: Gallimard.

Benjamin, W. (1989), Paris, capitale du XIXe siècle – le livre des passages, Paris: Le Cerf.

Berque, A. (2000), Écoumène, introduction à l'étude des milieux humains, Paris: Belin.

Bonfiglioli, S., Mareggi, M. (dir.) (1997), Il tempo e la città fra natura e storia. Atlante di progetti sui tempi della città, Urbanistica Quaderni, Rome : Inu Edizioni, n. 12.

Bonfiglioli, S., et Zedda, R. dir. (1999), Il piano dei tempi e degli orari di Pesaro, *Urbanistica Quaderni*, n. 18.

Cauvin, C. (1994), Accessibilité de système et accessibilité locale, *Révue Flux*, 10(16) : 39-49.

De Biase, A. (2014), Hériter de la ville. Paris: Donner lieu.

Deleuze, G., Guattari, F. (1980), Mille Plateaux, Paris: Minuit.

Foucault, M. (1984), Dits et écrits : Des espaces autres, Conférence au Cercle d'études architecturales, 14 mars 1967, in *Architecture, Mouvement, Continuité*, n°5, pp. 46-49.

Guez, A., Biase de, A., Gatta, F., Zanini, P. (2016), Chronotopic exploration of a parisian landscape, in *Temporalités de la ville*, à paraître aux éditions Europia.

Guez, A., Mareggi, M. (1997), Représentations de temporalités urbaines, in S. Bonfiglioli, M. Mareggi (dir.), *Il tempo e la città fra natura e storia. Atlante di progetti sui tempi della città, Quaderni di Urbanistica*, Rome: Inu Edizioni.

Guez, A., Stabilini, S., Zedda, R. (2000), Les temps italiens se réorganisent, in *La Recherche*, N° 337, supplément ville.com, décembre 2000.

Hägerstrand, T. (1970), What about people in regional science ? : *Papers of the Regional Science Association*, 24 : 6– 21.

Lefebvre, H. (1992), Éléments de rythmanalyse, Paris: Syllepse.

Louppe L, Dobbels D., Virilio P., Thom R., Laurenti J.-N., Dunlop V. (1994), Danses Tracées - Dessins et Notations des Chorégraphes. Paris: Dis Voir.

Lynch, K. (1972), What time is this place ? Cambridge : MIT Press.

Martinotti, G. (1993), Metropoli. La nuova morfologia sociale della città, Bologna : Il Mulino.

Piaget, J. (2012), La psychologie de l'enfant, Paris: PUF.

Piaget, J. Inhelder, B. (1959), La genèse des structures logiques élémentaires, Delachaux & Niestlé.

Rousseaux, F. (2006), La collection, un lieu privilégié pour penser ensemble singularité et synthèse, *revue EspacesTemps.net*: http://www.espacestemps.net/articles/la-collection-un-lieu-privilegie-pour-penser-ensemble-singularite-et-synthese/

Rousseaux, F., Saurel, P., Petit, J. (2014), Knowledge Engineering or Digital Humanities? Territorial Intelligence, a Case in Point. Innovation in Intelligent Machines-4, *Recent Advances in Knowledge Engineering: Paradigms and Applications*, Springer-Verlag, Vol. 514, pp. 129-187, C. Faucher and L-C. Jain eds.

Rousseaux, F., Legrand, J., Soulier, E., Bugeaud, F., Saurel, P., Neffati, H. (2012), A New Methodology for Collecting and Exploiting Vast Amounts of Dynamic Data, *3rd International Conference on Emerging Intelligent Data and Web Technologies*, 2012.

Rousseaux, F., Soulier, E., Saurel, P., Neffati, H. (2012), Agencement multi-échelle de territoires à valeur ajoutée numérique : des Hétérotopies foucaldiennes aux Complexes simpliciaux, Politiques publiques, Systèmes complexes, Edité par Danièle Bourcier, Romain Boulet et Pierre Mazzega, Editions Hermann, pp. 169-192, Paris.

Van Schaick, J. (2011), Timespace matters. Exploring the gap between knowing about activity patterns of people and knowing how to design and plan urban areas and regions, Delft: Eburon Academic Publishers. URL : http://www.eburon.nl/timespace_matters.

Wajcman, G. (1999), Collection, Paris: Nous.

# Cognisearch Business : un service de recherche d'information d'entreprises sur le web

Armel Fotsoh Tawofaing[*,**]
Annig Le Parc-Lacayrelle[*]
Tanguy Moal[**]

[*]Laboratoire LUIPPA, BP-1155, 64013 PAU Université Cedex, France
aftawofaing@univ-pau.fr
annig.lacayrelle@univ-pau.fr
[**]Cogniteev, 2 Rue Doyen Georges Brus, 33600 Pessac, France
armel@cogniteev.com
tanguy@cogniteev.com

**Résumé.** Ce papier propose un service de recherche d'entreprises sur le web. Le besoin d'information est exprimé par des critères thématiques (activités et métiers de l'entreprise, produits commercialisés, ...) ainsi que des critères spatiaux (adresses). Le système comprend un premier module de filtrage de sites web d'entreprises. Un second module analyse ces sites afin d'en extraire automatiquement toutes les informations contextuelles (activités, métiers, produits, contacts, adresses, ...). Notre proposition se distingue des services similaires que l'on peut trouver sur le web par (i) la richesse des informations thématiques (catégorisation détaillée des activités, des produits, des métiers) ; (ii) la provenance web des informations ; (iii) la combinaison de critères de recherche thématique, spatial et plein texte. Un prototype est développé pour mettre en œuvre notre proposition.

## 1 Introduction

D'après la théorie de Longley et al. (2010), l'information géographique comporte trois composantes : la composante spatiale, la composante temporelle et la composante thématique. Dans les documents textuels, la composante thématique est traitée selon deux principales approches : soit en s'appuyant sur les mots clés (Wong et al. (1987)) soit en utilisant la sémantique contenue dans le texte (Fernandez et al. (2011)).

De plus en plus d'entreprises sont présentes sur le web et publient de nombreuses informations relatives à leurs activités, leurs métiers, leurs produits et leurs coordonnées (adresses, numéros de téléphone, numéros de fax, adresses mail). Partant de ce constat, le projet Cognisearch Business vise à exploiter ces données pour offrir un service de recherche d'entreprises capable de répondre à des besoins d'information du type « charpente en chêne au sud de Poyartin ». Ce type de recherche comporte une dimension thématique (charpente en chêne) et une dimension spatiale (au sud de Poyartin).

Plusieurs solutions sur le web proposent des services permettant de répondre à ce type de recherche. Nous citerons par exemple les Pages Jaunes [1] ou même Google Maps [2]. Cependant, ces solutions (i) ne permettent pas de prendre en compte toutes les relations topologiques dans la dimension spatiale ; (ii) peuvent mal interpréter la dimension thématique dans certains cas ; (iii) s'appuient en grande partie sur des données saisies manuellement.

Notre objectif est donc de construire un service de recherche d'information dédié aux entreprises qui s'appuie sur des données publiées sur le web par les entreprises et combine des critères spatiaux (prenant en compte différentes relations topologiques) et des critères thématiques avec la recherche plein texte. En ce qui concerne le spatial, une des difficultés majeurs est l'extraction des adresses. En effet, très peu de sites web utilisent les microformats ou des balises particulières pour les définir. Les adresses peuvent être situées n'importe où dans la page et les différentes informations qui les composent ne sont pas écrites forcément dans le même ordre. Ce service de recherche d'informations vise deux catégories d'utilisateurs : (i) les « particuliers », qui interagissent avec le service par des requêtes exprimées en langage naturel ; ii) les « organisations », dont l'objectif est d'avoir des données d'entreprises pour alimenter leurs propres services.

La suite de l'article est structurée comme suit : la section 2 présente les travaux connexes à notre proposition. La section 3 détaille l'approche proposée pour construire notre service, ainsi que l'architecture retenue pour sa réalisation. La section 4 présente, quant à elle, un prototype du service et des résultats expérimentaux. La section 5 conclut l'article et présente les perspectives envisagées.

## 2 Travaux connexes

Les systèmes de recherche d'information « classiques » comme Google, Bing ou même Yahoo, ne sont pas toujours efficaces lorsque le besoin d'information comprend des critères spatiaux et thématiques propres à un domaine. C'est ainsi que des systèmes spécialisés ont été développés pour répondre à des besoins d'informations contextualisés. Concernant la recherche d'entreprises, Triou et al. (2007) proposent un système de recherche sémantique et géolocalisé dédié aux entreprises ; les informations relatives à chaque entreprise sont renseignées manuellement et stockées dans une ontologie.

Ahlers (2013) propose également un système d'analyse de pages web, pour enrichir la base de données d'un fournisseur des Pages Jaunes et par là améliorer les résultats de la recherche. Un démonstrateur de recherche d'information (RI) s'appuyant sur la base ainsi enrichie est mis en œuvre. L'approche proposée consiste en effet à croiser les données des Pages jaunes, avec celles de l'annuaire DMOZ [3] pour identifier les pages web associées à une entreprise et en extraire des informations (adresses, numéros de téléphone, adresses mail, données commerciales, données fiscales). L'objectif de cette démarche est d'enrichir la base de données de l'annuaire à partir d'informations consolidées provenant de sources multiples.

Cependant, la proportion d'entreprises existantes dans une région et qui sont référencées sur DMOZ reste faible pour le cas de la France (par exemple, on a 2580 entrées DMOZ au

1. http ://www.pagesjaunes.fr/
2. http ://www.google.fr
3. http ://www.dmoz.org/

total pour la région d'Aquitaine alors qu'il y a plus de 250000 entreprises déclarées au registre du commerce), d'où l'enrichissement limité de la base de données.

Il existe aussi sur le web plusieurs services dédiés à la recherche d'informations d'entreprises. Nous pouvons les classer en trois catégories principales : (i) les fournisseurs de données, comme Factual [4] ou Axciom [5], qui collectent et commercialisent les informations d'entreprises. Les informations alimentant ces services proviennent généralement des «données ouvertes», du crawl du web ou même de plateformes partenaires ; (ii) les annuaires, qui sont des bases de données d'informations d'entreprises consultables en ligne. Dans cette catégorie, nous pouvons citer des services comme Google Maps, Google My Business [6], Pages Jaunes, ou même société.com (annuaire d'entreprises françaises déclarées au registre du commerce) ; (iii) les réseaux sociaux, qui sont beaucoup plus orientés partage d'informations et d'appréciations portant sur les commerces, les places et les événements d'une région. C'est le cas des services comme Yelp [7], Foursquare [8], Facebook Places [9]. Dans les catégories (ii) et (iii) les données proviennent, en général, de contributions (salariés et utilisateurs) et de fournisseurs de données.

Tous ces services ne vont pas assez loin dans l'interprétation des relations spatiales exprimées dans les besoins d'informations. En effet, les relations du type « au sud de ... » ne sont pas toujours interprétées et prises en compte lors de la recherche. Cependant Vaid et al. (2005) présentent une approche qui prend en compte ces détails dans l'interprétation de la partie spatiale du besoin d'information.

Deux hiérarchies d'organisation des activités et des produits ont été définies par l'INSEE [10] : la Nomenclature Française des Activités (NAF) et la Classification des Produits Française (CPF) respectivement. Pôle Emploi a également constitué une hiérarchie (Répertoire Opérationnel des Métiers et des Emplois : ROME) qui organise les métiers et emplois en catégories socio-professionnelles. Ces différentes hiérarchies sont utilisées dans notre approche pour la construction des bases de connaissances relatives aux activités, aux métiers et aux produits. Concernant la représentation des entités entreprises, le répertoire SIRENE (Système Informatique pour le Répertoire des ENtreprises et des Etablissements) de l'INSEE propose un modèle sur lequel s'appuie societe.com.

De nombreux travaux traitent de l'annotation d'information dans le texte : le processus d'annotation peut se faire de façon manuelle, semi-automatique ou automatique. Deux principales approches sont utilisées pour l'annotation automatique d'adresses dans le texte : une première utilise des patrons d'extraction (Borges et al. (2007), Blohm (2011)) et une deuxième s'appuie sur les techniques d'apprentissage (Loos et Biemann (2008), Taghva et al. (2005)). D'autre part, l'exploitation de la sémantique est au centre de plusieurs travaux de recherche relatifs à l'annotation du thème dans le texte, ceci pour différents domaines comme le droit (Mimouni et al. (2015)), le sport (Soner et al. (2012)), la médecine (Drame (2014)) ou encore l'agriculture (Roussey et Bernard (2015)). En effet, la formalisation de la connaissance d'un domaine précis selon une structure organisée, comme les ontologies, permet d'annoter

4. http ://www.factual.com/
5. http ://www.acxiom.fr/
6. http ://www.google.com/business/
7. http ://www.yelp.fr/
8. http ://fr.foursquare.com/
9. http ://fr-fr.facebook.com/places/
10. http ://www.insee.fr/fr/

des concepts (Nešić et al. (2010)) ou des relations sémantiques (Royer et al. (2015)) dans le texte.

# 3 Proposition

Nous proposons un service de recherche géolocalisé d'information d'entreprises alimenté en informations uniquement par les données publiques et les informations extraites du web. Notre proposition se distingue des services existants présentés dans l'état de l'art par son indépendance vis à vis des données enregistrées manuellement. Ce service se veut également plus précis dans l'interprétation de la zone spatiale couverte par les besoins d'informations en prenant en compte les différentes relations topographiques exprimées dans les requêtes. Comparativement aux autres travaux exploitant le web pour extraire les informations d'entreprises comme ceux de Ahlers (2013), dans notre cas, les sites web d'entreprises à partir desquels se fait l'extraction d'information sont identifiés par une heuristique à partir des données du registre du commerce. Cette démarche permet d'avoir un corpus plus exhaustif que celui construit à partir de l'annuaire DMOZ. Par ailleurs, au delà des données de localisation (adresses), nous extrayons également les métiers, produits et activités d'entreprises.

Notre contribution propose l'utilisation des hiérarchies NAF, CPF et ROME, pour organiser les connaissances relatives respectivement aux activités, aux produits et aux métiers des entreprises et annoter automatiquement ces informations dans les sites web. Nous proposons également une méthode basée sur les règles de grammaire pour annoter automatiquement les coordonnées d'entreprises (adresses, numéros de téléphone, numéros de fax et adresses mails).

La figure 1 décrit l'architecture de notre service de recherche d'entreprises sur le web :

— une première étape (figure 1.1) filtre les sites dédiés aux entreprises sur le web à partir de leurs données d'immatriculation ;
— une étape d'annotation (figure 1.3) permet d'extraire automatiquement les informations relatives aux activités, aux produits, aux métiers ainsi qu'aux coordonnées de chaque entreprise (adresses, numéros de téléphones, emails, fax) .
— les informations annotées viennent compléter les données d'immatriculation des entreprises afin de constituer un premier index d'informations d'entreprises. En parallèle, le texte des sites web est indexé afin de permettre la recherche plein-texte (figure 1.6).

Cette partie traite des modèles choisis, des principales bases de connaissances utilisées, avant de détailler chaque phase de la chaîne de traitement mise en œuvre.

## 3.1 Modèles de représentation

Dans la littérature, de nombreux modèles sont proposés pour la représentation de différents types d'entités. Le web sémantique définit par exemple des modèles comme FOAF pour la représentation des entités de type organisations ou personnes. De même, les microformats et microdonnées définissent également des modèles de représentation d'entités de types adresses, entreprises ou même produits. Cependant, ces modèles sont assez généraux et adaptés pour le partage de l'information sur le web.

Nous proposons un nouveau modèle de représentation d'entreprises qui reprend celui du répertoire SIRENE utilisé par société.com, et le complète avec les informations provenant des hiérarchies NAF, CPF et ROME relatives aux activités, produits et métiers d'entreprises

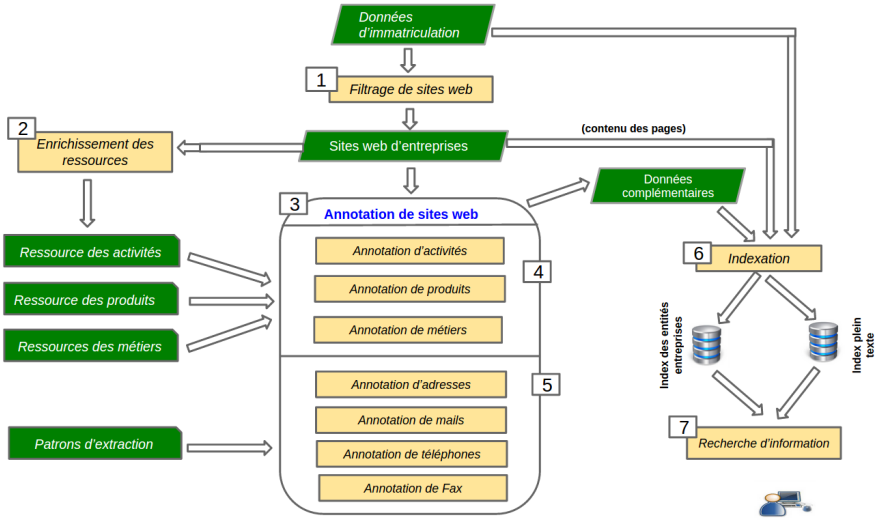



FIG. 1 – *Chaîne de traitement de Cognisearch Business*

respectivement. De même, les travaux de l'Etalab en collaboration avec l'IGN et la Poste ont permis de définir un modèle pivot de représentation des adresses françaises contenues dans la base nationale d'adresses [11]. Le modèle d'adresses choisi pour notre service est basé sur ce dernier.

Le tableau 1 détaille les principales caractéristiques de notre modèle de représentation d'entité entreprise. Pour chaque caractéristique, elle spécifie le modèle de référence utilisé, ainsi que la source d'information mobilisée pour construire une occurrence de l'entité.

## 3.2 Ressources pour l'annotation de sites web

Afin d'annoter automatiquement les métiers, les produits et les activités contenues dans les sites web, nous avons construit, par transformation de modèles, trois ressources de type connaissance (sous forme d'ontologies au format OWL), ceci à partir des hiérarchies NAF, CPF et ROME (cf. Travaux connexes). Elles décrivent respectivement les activités, les produits et les métiers d'entreprises. Le choix de ces ressources se justifie notamment par le niveau de finesse dans la description des différentes catégories.

L'ontologie relative aux activités ainsi construite est pauvre en vocabulaire. Nous avons de ce fait, mis en œuvre un processus pour son enrichissement de façon semi-automatique. Ce processus utilise l'apprentissage pour regrouper les expressions contenues dans les sites web, qui sont communes à une catégorie précise d'activité (figure 1.2). Les données d'immatriculation d'une entreprise référencent son activité principale via son code APE. De ce fait, à chaque classe de la hiérarchie des activités, on peut associer un ensemble de sites web d'entreprises. Un algorithme de clustering permet de mettre en évidence les expressions communes à tous

11. http ://adresse.data.gouv.fr

| **Entité Entreprise** | **Modèle de représentation** | **Source d'informations** |
|---|---|---|
| **Données d'immatriculation** | SIRENE (INSEE) | société.com |
| **Coordonnées** | | |
| Adresse site web | - | web |
| Adresses | Adresse (Etalab) | site web de l'entreprise |
| Téléphones, mails, fax | - | site web de l'entreprise |
| **Informations complémentaires** | | |
| Métiers | ROME (Pôle Emploi) | site web de l'entreprise |
| Activités | NAF (INSEE) | site web de l'entreprise |
| Produits | CPF (INSEE) | site web de l'entreprise |

TAB. 1 – *Modèle simplifié d'une entité entreprise*

les sites web de cette classe. Ces expressions constituent donc un vocabulaire potentiel pour enrichir la ressource. Une phase de validation par un expert intervient par la suite pour sélectionner les expressions pertinentes. L'algorithme de clustering que nous avons utilisé dans notre proposition est Latent Dirichlet Allocation (LDA), illustré dans Blei et al. (2003)

## 3.3 La chaîne de traitement

L'architecture proposée pour la mise en œuvre du service Cognisearch Business est constituée de plusieurs modules comme le montre la figure 1.

### 3.3.1 Constitution du Corpus

Les données d'immatriculation, telles que renseignées au registre du commerce sont issues de la ressource societe.com. Ces informations constituent les données de base d'une entité entreprise. Un traitement basé sur une heuristique exploite ces données d'immatriculation pour filtrer le web et retrouver, s'il existe, le site web associé à chaque entreprise (figure 1.1). L'ensemble des sites web ainsi obtenu, constitue les données d'entrées pour le processus d'enrichissement des ressources (figure 1.2) ainsi que pour le processus d'annotation (figure 1.3).

### 3.3.2 Extraction d'informations

Le contenu de chaque site web du corpus est ensuite analysé, afin d'extraire les informations complémentaires pour enrichir celles de base.

**Extraction d'informations thématiques**
L'extraction d'informations relatives aux métiers, aux produits et aux activités d'entreprises est réalisée en utilisant les ressources de type connaissance sous forme d'ontologies OWL (voir section 3.2). Les syntagmes nominaux correspondant aux vocabulaires associés à ces ressources sont identifiés dans le texte des pages web. Chacun de ces syntagmes est annoté de l'identifiant de la classe associée dans la ressource correspondante.

| **Noms de champs** | **Abréviations** | **Exemples** |
|---|---|---|
| Complément d'adresse | CA | Résidence Rigaud |
| Introducteur de Nom de Voie | INV | Avenue |
| Boite Postale | BP | BP 1167 |
| Course Spéciale | CS | CS 2587 |
| Numéro de Voie | NV | 10 ter |
| Nom de Voie | NVo | Avenue de l'université |
| Code Postal | CP | 64000 |
| Commune | C | Pau |
| Numéro de Courrier | NC | CEDEX 01 |
| Département | D | Pyrénées-Atlantiques |
| Pays | P | France |

TAB. 2 – *Champs de l'adresse*

En ce qui concerne les emails, les numéros de téléphone et numéros de fax, des patrons sont utilisées pour leur extraction. Ces patrons ont été mis au point à partir de l'observation d'un échantillon de sites web.

**Extraction d'information spatiales**

Notre objectif est d'extraire les adresses postales dans les sites web. Ahlers et Boll (2008) proposent une approche d'extraction d'adresses allemandes en utilisant plusieurs gazetiers, notamment un répertoriant les noms de toutes les voies. Dans le contexte français, il n'existe pas une ressource complète contenant tous les noms de voie. Une des difficultés est donc d'identifier le nom de la voie dans une adresse. En effet, il peut y avoir un ou plusieurs compléments qui peuvent être positionnés avant et/ou après le nom de la voie.

L'observation d'un échantillon de 160 sites web d'entreprises a permis d'identifier plusieurs formes d'expression pour les adresses françaises. Le tableau 2 répertorie les différentes informations constituant une adresse dans sa forme complète. Le code postal est le seul champ présent dans toutes les formes identifiées.

Voici un extrait des patrons écrits à partir de l'observation des 160 sites. La tableau 3 correspond à la légende utilisée pour écrire les patrons.

**Patron 1**

$$\textbf{Adresse} \rightarrow CA? \quad ((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad NV? \quad NVo \quad CA?$$
$$((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad ((CP\,C) \mid (C\,CP)) \quad NC? \quad D? \quad P?$$

**Patron 2**

$$\textbf{Adresse} \rightarrow CA? \quad ((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad ((CP\,C) \mid (C\,CP))$$
$$NC? \quad D? \quad P?$$

TAB. 3 – *Légende des patrons d'extraction*

| | |
|---|---|
| A ? | A est facultatif |
| A \| B | A ou B |
| A B | A suivi de B |

**Patron 3**

$$\textbf{Adresse} \rightarrow CA? \quad ((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad NV? \quad NVo \quad CA? \quad ((BP \quad CS) \mid (CS \quad BP) \mid BP \mid CS)? \quad CP \quad NC? \quad D? \quad P?$$

Dans le premier patron, le code postal, la commune et le nom de voie sont obligatoires, et peuvent être complétés par d'autres informations («10 Rue du Maréchal Foch, 49000 Angers», «Résidence des Aubiers, 3e Étage, 14 ter Rue de la République, 64000 Pau, France»). Cette forme est la plus fréquente dans l'échantillon (75% des adresses). Pour le deuxième patron, le code postal et la commune sont obligatoires et le nom de voie est absent («Résidence Rigaud 33350 Mouliets-et-Villemartin»). Le patron 3 permet entre autre d'identifier le cas où on a un code postal et un nom de voie sans commune («10 Place de la République, F-33600»). Cette dernière forme est assez rare (moins de 4% des adresses de l'échantillon).

### 3.3.3 Indexation

Les annotations sont extraites des pages web pour la construction des entités finales à indexer. Ces annotations sont rajoutées aux données d'immatriculation de chaque entreprise ainsi que l'adresse du site web correspondant. Les coordonnées GPS de chaque adresse extraite sont calculées avant l'ajout. L'entité finale respectant le modèle défini en 3.2 est stockée dans un index (figure 1.6). De plus, une opération parallèle consiste à extraire le contenu des pages de chaque site web d'entreprise et à l'indexer.

Les index ainsi construits sont utilisés pour répondre des besoins d'information qui supportent des critères d'interrogation multidimensionnels et exploitent les caractéristiques spatiales, thématiques et plein texte contenu dans les index (figure 1.7).

# 4 Prototype

Un prototype qui s'appuie sur l'architecture proposée dans la section précédente a été développé. Il implémente toutes les phases de la chaîne de traitement.

## 4.1 Corpus

Le prototype traite des données relatives aux entreprises de la région d'Aquitaine. Pour cette région, nous avons identifié 254 000 entreprises. Parmi ces entreprises, nous nous sommes intéressés uniquement à celles traitant de 6 domaines d'activités : commerce, construction,

hébergement & restauration, enseignement, information & communication et activités scientifiques & techniques. Ceci a réduit la liste initiale à 115 000 entreprises. Des sites web ont été automatiquement identifiés pour 22 000 d'entre elles. Le corpus d'analyse relatif à ces 22 000 entreprises est constitué de 550 000 pages web. C'est à partir de ce corpus que sont extraites les informations complémentaires pour la construction des entités entreprises à indexer.

## 4.2 Mise en œuvre de la chaîne de traitement

A partir des données d'immatriculation de société.com, une heuristique est utilisée pour le filtrage de sites web (figure 1.1). Ce processus s'appuie sur les caractéristiques nominatives de l'entreprise (nom commercial et nom officiel), ainsi que sur les propriétés de localisation (ville). Des requêtes sont construites et soumises de façon automatique à la plateforme google.fr via le framework de création de robots d'indexation Scrapy [12]. Ceci permet de générer la liste des 22 000 sites web d'entreprises. Cette liste est utilisée pour le téléchargement des pages web associées en vue de la constitution du corpus. C'est l'outil Nutch [13] de la fondation Apache qui est utilisé à cet effet.

En ce qui concerne l'enrichissement des bases de connaissances, nous avons utilisé la bibliothèque Python lda [14] qui implémente l'algorithme LDA conformément à l'approche décrite dans la section 3.2. Ce processus a permis d'extraire du vocabulaire à associer aux classes de la ressources des activités, afin d'améliorer l'annotation du corpus.

La phase d'annotation du corpus est mise en œuvre en utilisant l'API JAVA du framework GATE [15]. En effet, GATE a un module qui permet d'annoter des documents textuels en se basant sur une ressource de type ontologique au format OWL. De même, il est possible avec le module JAPE (Java Annotation Patterns Engine) de GATE, d'écrire avec une syntaxe propre, des règles d'extraction de patrons dans le texte. Ce sont ces deux modules que nous avons utilisés respectivement pour l'annotation des activités, des produits et des métiers dans un premier temps et des adresses, mails, téléphones et fax dans un second temps. En outre, pour le traitement de la langue française, nous avons intégré TreeTagger [16] à GATE. Des gazetiers contenant notamment l'ensemble des communes de France, les types introducteurs de voies (rue, avenue ...) sont également utilisés pour la détection d'adresses. Pour des raisons techniques et vu la volumétrie du corpus, nous avons mis en œuvre le processus d'annotation en utilisant la plateforme HADOOP [17], afin de gérer de façon distribuée et efficace cette étape de la chaîne de traitement.

Les entités entreprises, construites à partir des annotations extraites, sont indexées sous Elasticsearch [18]. Le contenu textuel des pages du site web de chaque entreprise est également indexé. Les deux index ainsi construits (index des entités entreprises et index plein texte) ont une taille globale de l'ordre de 3 GB.

L'interrogation du service, s'appuyant sur les deux index construits précédemment, se fait via l'interface web du plugin MARVEL [19] d'Elasticsearch. Les requêtes sont écrites en utilisant

12. http ://scrapy.org/
13. http ://nutch.apache.org/
14. http ://pypi.python.org/pypi/lda
15. http ://gate.ac.uk/
16. http ://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/
17. http ://hadoop.apache.org/
18. http ://www.elastic.co/
19. http ://www.elastic.co/products/marvel

la syntaxe du DSL (Domain Specific Language) d'Elasticsearch et les documents classés par ordre de pertinence sont retournés au format JSON.

## 4.3 Expérimentation du prototype

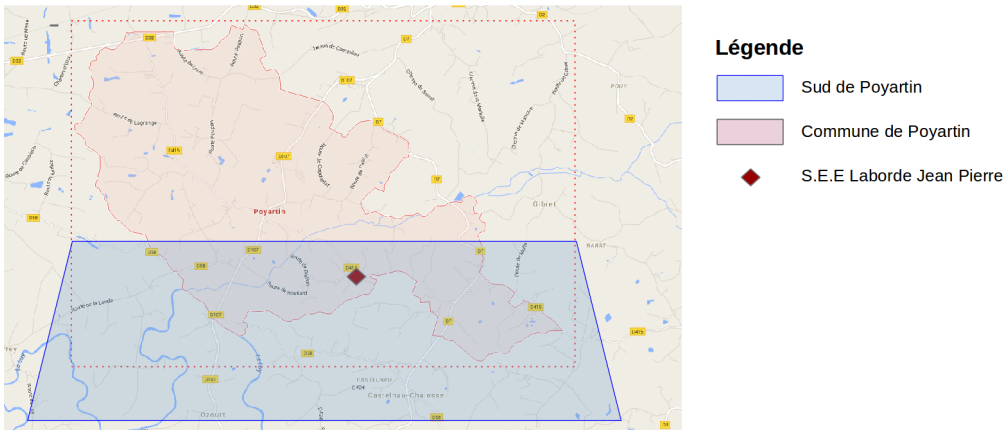



FIG. 2 – *Zone spatiale de recherche*

Le modèle de recherche d'information utilisé dans cette version du prototype est celui embarqué par défaut dans le moteur Lucene [20] d'Elasticsearch. Il s'agit en fait du modèle vectoriel (TF-IDF). D'autres modèles comme le modèle probabiliste (BM25) sont également disponibles.

La mise en oeuvre de la chaîne de traitement sur notre corpus de 22 000 sites web d'entreprises a permis d'annoter 30 000 adresses, 44 000 labels d'activités, 12 500 labels de produits et 28 000 labels de métiers.

Nous avons interrogé les deux index construits avec la requête suivante : « charpente en chêne au sud de Poyartin ».

— L'expression « charpente en chêne » définit l'empreinte thématique de la requête. Elle est identifiée, en se basant sur la ressource des activités comme un label de la classe d'activités « travaux de charpente ».
— L'expression « au sud de Poyartin » définit l'empreinte spatiale de la requête. La commune de Poyartin est représentée sur la figure 2 par la région limitée par les courbe en rouge. L'approche proposée par Sallaberry et al. (2008) permet d'approximer le « sud de Poyartin » par la zone trapézoïdale qui couvre la partie australe de la commune et ses abords (région en bleu ciel de la figure 2).

Les deux dimensions ainsi identifiées dans la requête sont soumises au moteur de RI d'Elasticsearch avec des poids identiques. Le système Cognisearch Business retourne, pour ce besoin d'information, une liste d'entreprises. Les informations de la première entreprise de cette liste

20. http ://lucene.apache.org/

| **Propriétés** | **Valeurs correspondantes** |
|---|---|
| Nom d'immatriculation | S.E.E. LABORDE JEAN PIERRE |
| Nom Commercial | CONCEPT CHALOSSE SOLAIRE |
| Numéro de SIRET | 43204808000019 |
| RCS | Dax B 432 048 080 |
| Statut juridique | Société à responsabilité limitée |
| Capital Social | 8000 € |
| Manager(s) | M. Jean-Pierre LABORDE |
| Adresse(s) | 1323 Route Abbadie, 40380 POYARTIN |
| Numéro(s) de téléphone | 05.58.98.69.19 |
| Mail(s) | contact@see-laborde.fr |
| Site web | http ://www.see-laborde.fr |

TAB. 4 – *Exemple de résultat*

sont présentées dans le tableau 4. Le site web de cette entreprise présente bien l'activité « travaux de charpentes » parmi des activités principales, et on peut vérifier sur la figure 2 qu'elle est bien dans la région spécifiée par le besoin d'information.

La même requête soumise à Google Maps (figure 3.1), ne retourne aucun résultat, sûrement parce que la requête exprimée n'a pas pu être interprétée par le système. Par ailleurs, la précision de localisation « au sud de Poyartin » ne peut être exprimée avec les Pages Jaunes, la seule relation topologique proposée étant « à proximité ». Si l'on lui soumet la requête « charpentes en chêne à proximité de Poyartin », on constate que la dimension thématique est mal interprétée car le premier résultat retourné est un restaurant (figure 3.2).

# 5 Conclusion

Nous avons présenté dans cet article un nouveau service de recherche géo-localisé d'informations portant sur les entreprises. Ce service s'appuie sur plusieurs problématiques de recherche différentes, qui deviennent complémentaires dans le processus de construction des entités entreprises. Il s'agit notamment de l'apprentissage qui est utilisé pour l'enrichissement des ressources de type connaissance. De plus, le processus d'annotation de texte combine les approches basées sur les patrons d'extraction et celles exploitant les bases de connaissances. Il permet d'extraire les adresses des sites malgré les différentes formes que l'on peut trouver. Un nouveau modèle de représentation d'entité entreprise est également proposé. Un premier démonstrateur est élaboré pour mettre en œuvre l'architecture de l'approche que nous avons définie. Ce prototype démontre la faisabilité et l'intérêt de l'approche. Il valide la chaîne de traitement et illustre le potentiel du système à traiter des requêtes «complexes». Une évaluation du processus d'extraction d'information ainsi que du filtrage des sites web d'entreprises est en cours de réalisation.

L'article décrit la première étape de la construction du service, qui est centrée sur l'extraction des informations sur le web et leur restructuration dans des index. Une deuxième étape consistera à exploiter ces index pour répondre à des besoins d'information combinant les cri-

tères spatiaux, thématiques et plein texte. Une évaluation du service avec un jeu de requêtes représentatif sera également menée au terme de cette étape.

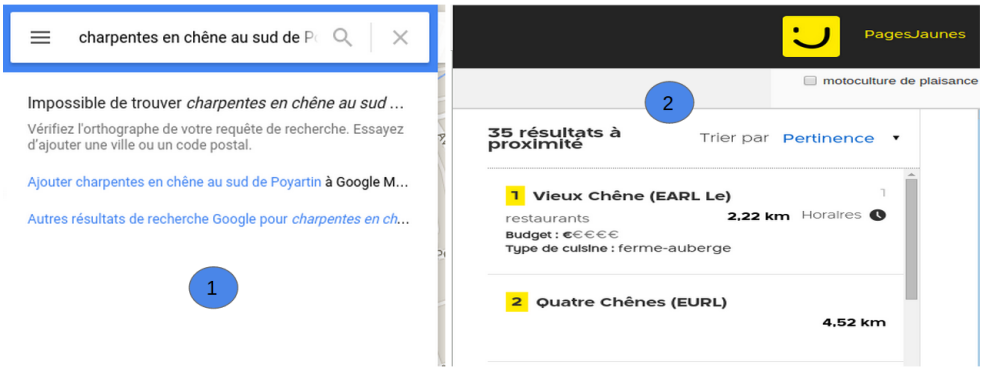



FIG. 3 – *Expérimentation Page Jaunes & Google Maps*

## Références

Ahlers, D. (2013). Business entity retrieval and data provision for yellow pages by local search. In *IRPS Workshop@ ECIR2013*.

Ahlers, D. et S. Boll (2008). Retrieving address-based locations from the web. In *Proceedings of the 2Nd International Workshop on Geographic Information Retrieval*, GIR '08, New York, NY, USA, pp. 27–34. ACM.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *the Journal of machine Learning research 3*, 993–1022.

Blohm, S. (2011). *Large-scale pattern-based information extraction from the world wide web*. KIT Scientific Publishing.

Borges, K. A. V., A. H. F. Laender, C. B. Medeiros, et C. A. Davis, Jr. (2007). Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, GIR '07, pp. 31–36. ACM.

Drame, K. (2014). *Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical*. Ph. D. thesis. Thèse de doctorat dirigée par Salamon Roger, Diallo Gayo et Mougin, Fleur Santé publique. Option Informatique et Santé.

Fernandez, M., I. Cantador, V. Lopez, D. Vallet, P. Castells, et E. Motta (2011). Semantically enhanced information retrieval : An ontology-based approach. *Web Semantics : Science, Services and Agents on the World Wide Web 9*(4), 434 – 452.

Longley, P. A., M. Goodchild, D. J. Maguire, et D. W. Rhind (2010). *Geographic Information Systems and Science* (3rd ed.). Wiley Publishing.

Loos, B. et C. Biemann (2008). supporting web-based address extraction with unsupervised tagging. In *Data Analysis, Machine Learning and Applications 2008*, pp. 577–584.

Mimouni, N., A. Nazarenko, et S. Salotti (2015). Vers une recherche sémantique et à base de graphe dans les systèmes d'accès à l'information juridique. In *Actes RISE 2015*, pp. 18–29.

Nešić, S., F. Crestani, M. Jazayeri, et D. Gašević (2010). Concept-based semantic annotation, indexing and retrieval of office-like document units. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, Paris, France, France, pp. 134–135.

Roussey, C. et S. Bernard (2015). Annotation des bulletins de santé du végétal. In *Actes RISE 2015*, pp. 43–54.

Royer, A., C. Sallaberry, A. Le Parc-Lacayrelle, et M.-N. Bessagnet (2015). Extraction automatique de relations sémantiques définies dans une ontologie. In *Actes RISE 2015*, pp. 30–42.

Sallaberry, C., M. Gaio, D. Palacio, et J. Lesbegueries (2008). Fuzzying gis topological functions for gir needs. In *Proceedings of the 2Nd International Workshop on Geographic Information Retrieval*, GIR '08, New York, NY, USA, pp. 1–8. ACM.

Soner, K., A. Ozgur, S. Orkunt, A. Samet, C. N. K., et A. F. N. (2012). An ontology-based retrieval system using semantic indexing. Volume 37, Oxford, UK, UK, pp. 294–305.

Taghva, K., J. S. Coombs, R. Pereda, et T. A. Nartker (2005). Address extraction using hidden markov models. In *Proceedings of IST/SPIE 2005 Int Symposium on Electronic Imaging Science and Technology, San Jose, California*, pp. 119–126.

Triou, F., F. Picarougne, et H. Briand (2007). Apport du web sémantique dans la réalisation d'un moteur de recherche géo-localisé à usage des entreprises. In *Extraction et gestion des connaissances (EGC'2007), Actes des cinquièmes journées Extraction et Gestion des Connaissances, Namur, Belgique, 23-26 janvier 2007, 2 Volumes*, pp. 69–80.

Vaid, S., C. B. Jones, H. Joho, et M. Sanderson (2005). Spatio-textual indexing for geographical search on the web. In *Proceedings of the 9th International Conference on Advances in Spatial and Temporal Databases*, SSTD'05, Berlin, Heidelberg, pp. 218–235.

Wong, S. K., W. Ziarko, V. V. Raghavan, et P. C. Wong (1987). On modeling of information retrieval concepts in vector spaces. *ACM Trans. Database Syst. 12*(2), 299–321.

## Summary

Searching information about local businesses is not a trivial problem to address. Most of existing services are supplied with manually recorded data. Based on the observation that more and more businesses are presented on the web, we propose in this paper a new approach, which consists to extract companies targeted information (addresses, activities, jobs, products, emails, fax) from websites, to supply a local businesses search service. The retrieval information module combines thematic, spatial and full-text criteria. A prototype of this service is implemented to experiment our proposal.

# La confiance est dans l'air ! Application à l'identification des parcours hospitaliers

Yves Mercadier*, Jessica Pinaire*,**,***
Jérôme Azé*,Sandra Bringay*,**** Maguelonne Teisseire*,♯

* LIRMM, UMR 5506, Université Montpellier, France
prenom.nom@lirmm.fr,
** CHU, Département d'information médicale, BESPIM, Nîmes, France
jessica.pinaire@chu-nimes.fr
*** équipe d'accueil 2415, Institut Universitaire de Recherche Clinique,
Université Montpellier, Montpellier, France
paul.landais@umontpellier.fr
**** AMIS, Université Paul Valéry, Montpellier, France
Sandra.Bringay@univ-montp3.fr
♯ TETIS, IRSTEA, Montpellier, France
maguelonne.teisseire@irstea.fr

**Résumé.** L'extraction de motifs séquentiels permet d'identifier les séquences fréquentes d'événements ordonnés. Afin de résoudre le problème du grand nombre de motifs obtenus, nous proposons l'extension pour les motifs séquentiels de la confiance, mesure d'intérêt utilisée classiquement pour sélectionner les règles d'association. Dans cet article, après avoir présenté les données, nous définirons formellement la notion de confiance appliquée aux motifs séquentiels. Nous appliquerons cette mesure pour identifier des trajectoires hospitalières, représentées par les motifs séquentiels, dans des données issues du PMSI (Programme de Médicalisation des Systèmes d'Information). Nous nous sommes focalisés sur un cas d'étude hospitalière : l'infarctus du myocarde (IM), et notamment la prédiction de la trajectoire des patients ayant eu un IM entre 2009 et 2013. Les résultats obtenus ont été soumis à un spécialiste pour discussion et validation.

## 1 Introduction

Parmi les méthodes d'extraction de connaissances non supervisées, nous nous intéressons aux méthodes d'extraction de motifs. Il en existe un très grand nombre permettant d'identifier des régularités dans les jeux de données. L'ingénieur de la connaissance utilise alors son expérience pour proposer un type ou une combinaison de ces motifs selon les besoins des experts métier souhaitant exploiter le jeu de données. Ces méthodes génèrent très souvent un grand nombre de motifs. Un expert métier peut alors être submergé d'informations et ne pas tirer partie des résultats du processus de fouille de données. Pour limiter le nombre de motifs présentés et permettre leur interprétation, l'ingénieur de la connaissance applique alors un filtrage

des motifs en appliquant des mesures d'intérêt puis regroupe les motifs en appliquant des mesures de similarité. Les mesures d'intérêt sont des indicateurs statistiques dont la sémantique est parfois difficile à interpréter par l'expert.

De nombreuses études ont été réalisées en vue de comparer les mesures d'intérêt pour aller vers une amélioration du résultat final de la fouille de données par exemple Lenca et al. (2003), Blanchard (2005b). Dans le cadre de cette étude, nous nous sommes intéressés à une mesure en particulier, la confiance. Cette mesure d'intérêt a été introduite par Agrawal et al. (1993) pour les règles d'associations. Elle consiste à estimer la probabilité dans la base de données d'obtenir une association à partir de ses constituants. L'originalité de notre mesure que nous avons appelée r-confiance est double. Premièrement, elle fonctionne pour tous les types de motifs (règle d'association, motif séquentiel, motif spatio-temporel). Deuxièmement, elle utilise comme opérateur d'agrégation « la proportion de position ». Nous avons également développé une interface qui permet de représenter les motifs extraits mais également de les comparer en prenant en compte différentes mesures d'intérêt, dont la r-confiance.

Nous avons appliqué cette mesure dans un contexte spécifique pour identifier des trajectoires hospitalières, représentées par des motifs séquentiels, dans des données issues du PMSI (Programme de Médicalisation des Systèmes d'Information). Nous nous sommes focalisés sur un cas d'étude hospitalière : l'infarctus du myocarde (IM) et notamment la prédiction de la trajectoire des patients ayant eu un IM entre 2009 et 2013. Nous cherchons à identifier des trajectoires de GHM (Groupe Homogène de Malade), ce dernier est un code renseignant les caractéristiques d'une hospitalisation. Les résultats obtenus ont été soumis à un spécialiste pour discussion et validation.

Dans cet article, nous allons présenter la r-confiance dans la section 2 ainsi que l'interface de l'outil développé dans la section 3. Nous présenterons notre cas d'étude et les données utilisées dans la section 4. Nous analyserons les résultats obtenus et démontrerons l'efficacité de cette méthode, que nous pourrons utiliser à plus grande échelle, par exemple, dans l'identification de trajectoires fréquentes de patients pour un contexte donné.

# 2 Sélection des motifs d'intérêt selon la r-confiance

## 2.1 Vers une nouvelle confiance

Depuis les années 90, de nombreuses méthodes ont été proposées pour l'extraction de motifs fréquents dans les bases de données Rabatel (2011). Ces motifs se sont complexifiés avec le temps pour prendre en compte différentes dimensions (temporelles, spatiales...)(Heas, 2005). On peut citer les règles d'association (Agrawal et al., 1996), les motifs séquentiels (Pei et al., 2001), les co-localisations (Sundaram et al., 2012), les trajectoires (Etienne et Devogele, 2012), les graphes (Pennerath et Napoli, 2006). Par ailleurs, afin de résoudre le problème du grand nombre de règles d'association de nombreuses mesures d'intérêt ont été proposées pour leur sélection (Tan et al., 2002). D'après Grissa (2013) on peut dénombrer plus de soixante mesures d'intérêt dédiées aux règles d'association. À notre connaissance il n'existe que très peu de mesure d'intérêt pour les motifs séquentiels, les co-localisations, les graphes ou encore les trajectoires. Nous avons voulu combler ce vide en proposant une mesure d'intérêt spécifique applicable à tout type de motif. Nous avons choisi de faire l'extension de la confiance définie pour les règles d'association car cette mesure permet d'estimer la liaison entre les deux itemsets

constituant l'association. Nous pensons que la qualité des liaisons inter-itemset dans un motif est primordiale pour estimer et discriminer les motifs dans un ensemble de motif.

## 2.2 Motifs séquentiels et candidats séquentiels

En 1995, la problématique de la recherche de règles d'association a été étendue pour détecter des comportements typiques dans le temps et le concept de motifs séquentiels a été introduit (Agrawal et Srikant, 1995) puis les règles séquentielles ont été proposées par Das et al. (1998). Ces motifs ont été appliqués dans de nombreux domaines comme le panier de la ménagère précédemment introduit (Agrawal et Srikant, 1995), la fouille de données d'usage du Web (Dong Haoyuan et al., 2009), la fouille de texte (Charnois et al., 2009).

Nous présentons quelques définitions préliminaires des motifs séquentiels avant de proposer notre mesure de filtrage.

**Définition 1.** *Une base de données séquentielles contient un ensemble ordonné d'éléments, généralement par le temps. Soit* $I = \{i_1, i_2, ..., i_k\}$ *l'ensemble de tous les items. Un sous-ensemble de* $I$ *est appelé un itemset. Une séquence* $S =< I_1, I_2, ..., I_n >$ *est une liste ordonnée d'itemset (*$I_i \subseteq I$*). Chaque itemset d'une séquence représente un ensemble d'événements qui apparaissent à la même estampille temporelle. Les différents itemsets d'une séquence sont associés à des estampilles temporelles différentes.*

**Définition 2.** *Une séquence* $S_1 =< I_1, I_2, ..., I_m >$ *est une sous-séquence de* $S_2 =< I'_1, I'_2, ..., I'_n >$ *(noté* $S_1 \preceq S_2$*) si et seulement si* $\exists i_1, i_2, ..., i_m$ *tels que* $1 < i_1 < i_2 < ... < i_m \leq i_n$ *et* $I_1 \subseteq I'_1, I_2 \subseteq I'_2, ..., I_m \subseteq I'n$.

**Exemple 1.** Un patient se rend dans un hôpital pour une pathologie : son séjour est codé par un item ex. code -> 01. Il revient plus tard pour un deuxième examen ex.code -> 02. Cela constitue une séquence dans la base que l'on peut noter : $S_0 =< (01), (02) >$ la séquence $S'_0 =< (01) >$ est une sous-séquence de $S_0$

**Définition 3.** *Etant donné un ensemble de séquences* $D = \{S_1, S_2, ..., S_n\}$*, le support d'une séquence* $S_1$ *correspond au nombre de séquences de* $D$ *qui contiennent* $S_1$*. Si le support d'une séquence* $S_1$ *satisfait un seuil de support minimum* $minsup$*, alors* $S_1$ *est une séquence fréquente ou motif séquentiel. L'objectif de la recherche de motifs séquentiels est donc d'extraire l'ensemble complet des séquences fréquentes par rapport à un seuil de support minimum minsup.*

**Exemple 2.** Un deuxième patient se rend dans un hôpital pour une série d'examen : sont séjour est codé par un item ex. code -> 03. Il revient plus tard pour une deuxième série d'examen ex.code -> 04. Un troisième patient fait de même. $S_1 =< (03), (04) > S_2 =< (03), (04) >$ Nous fixons arbitrairement le seuil de support à deux. La séquence $< (03), (04) >$ sera alors considérée fréquente.

**Définition 4.** *Etant donné un motif séquentiel* $M =< M_1, M_2, ..., M_n >$*, un* **candidat séquentiel** *de* $M$*,* $C_p =< C_1, C_2, ..., C_p >$*, est défini comme une des sous-séquences préfixes de* $p$ *items telle que* $p < m$ *et* $\forall i_1, i_2, ..., i_p$*,* $1 \leq i_1 < i_2 < ... < i_p$ *et* $C_1 = M_{i_1}, ..., C_{p-1} = M_{m-p-1}, C_p = M_{m-p}$*. Un motif séquentiel de longueur* $n$ *peut être associé à* $n-1$ *candidats séquentiels.*

Soit $M =< (ab)(c)(d)(e) >$ un motif. Ce motif génère 3 candidats séquentiels : $< (ab) >$, $< (ab)(c) >$ et $< (ab)(c)(d) >$. $< (ab)(c) >$ est un candidat séquentiel mais pas $< (ab)(d) >$ car ses itemsets ne sont pas consécutifs dans $M$.

## 2.3 La r-confiance

Nous proposons de considérer la confiance que l'on peut avoir dans un motif comme étant la mesure de la représentativité de ce motif par rapport aux données. Plus la confiance d'un motif est élevée, plus l'apparition des premiers items le constituant permet l'obtention du motif complet. Concrètement, la confiance exprime alors la qualité de la liaison des itemsets internes au candidat à partir du premier itemset le constituant. Avant de définir la r-confiance d'un motif de façon générale, nous définissons la r-confiance élémentaire d'un motif séquentiel.

Soit M un motif de longueur n et C un candidat séquentiel de ce motif de longueur p selon la définition 4. La r-confiance élémentaire, notée *r-conf-e*, est définie par :

$$\textit{r-conf-e}(M, C) = \frac{support_B(M)}{support_B(C)} \tag{1}$$

La r-confiance calculée pour le motif M correspond à l'agrégation des n-1 r-confiances élémentaires des candidats séquentiels le composant, c'est-à-dire la proportion de candidats séquentiel ayant une r-confiance élémentaire. Afin de conserver la notion de mesure d'intérêt et donc de filtrage des motifs extraits, seules les r-confiances élémentaires dont la valeur est supérieure à un seuil fixé minR seront prises en compte dans cette agrégation. Pour M un motif de longueur n, soit $\mathbf{C}$, l'ensemble des n-1 candidats séquentiels de M.

$$\textit{r-conf}(M) = \begin{cases} 0 & \text{si } Card(\{C \in \mathbf{C}, \textit{r-conf-e}(M, C) > minR\}) = 0 \\ \frac{Card(\{C \in \mathbf{C}, \textit{r-conf-e}(M,C) > minR\}) + 1}{n} & \text{sinon} \end{cases} \tag{2}$$

Cette mesure, ici définie pour les motifs séquentiels, est certainement généralisable à d'autres types de motif. Nous considérons ici que chaque motif est construit structurellement à partir de l'ensemble des candidats de celui-ci. En effet la r-confiance mesure une caractéristique de la structure du motif, elle évalue la proportion de candidat ayant une r-confiance élémentaire supérieure à un seuil. Les motifs de tous types pouvant être considérés comme des conteneurs structurés, il semble possible d'en évaluer la qualité par cette mesure.

## 2.4 Exemple de calcul

Considérons la base de données du tableau 1 constituée de sept motifs séquentiels.
Nous étudions le motif séquentiel $\mathbf{M} = \langle (a, b)(c)(b, c) \rangle$. Le calcul des supports donne :
$support(\langle (\mathbf{a}, \mathbf{b}) \rangle) = \frac{3}{7} = 0.43$
$support(\langle (\mathbf{a}, \mathbf{b})(\mathbf{c}) \rangle) = \frac{2}{7} = 0.29$

Nous en déduisons d'abord le calcul des r-confiances élémentaires :
*r-conf-e*$(\mathbf{M}, \langle (a, b) \rangle) = \frac{\frac{2}{7}}{\frac{3}{7}} = 0.66$
*r-conf-e*$(\mathbf{M}, \langle (a, b)(c) \rangle) = \frac{\frac{2}{7}}{\frac{2}{7}} = 1$

TAB. 1: Exemple d'une base de motifs séquentiels.

| Séquences |
|---|
| $\langle(a,b)(c)(b,c)\rangle$ |
| $\langle(c)(c)(b)\rangle$ |
| $\langle(d)(e)(f)\rangle$ |
| $\langle(d)(f)\rangle$ |
| $\langle(a,b)(d,e)\rangle$ |
| $\langle(d)(a,b)(c)(b)(b,c)\rangle$ |
| $\langle(c)(b)(d)(b,c)\rangle$ |

Puis, à partir d'un seuil fixé à $minR = 0.7$, nous pouvons calculer la r-confiance du motif : *r-conf*$(\mathbf{M}) = \frac{1+1}{3} = 0.66$

# 3 Présentation de l'interface utilisateur

La phase d'extraction des motifs conduit souvent à générer un trop gros volume de motifs à valider. L'expert est alors submergé et démuni devant les nouvelles données ainsi produites. Notre proposition d'une nouvelle mesure de filtrage s'inscrit dans ce contexte et a été implémentée dans une interface interactive avec pour objectif de faciliter l'analyse des résultats de la fouille de données séquentielles. Ce problème a déjà été approché pour le traitement de fouille de règle d'association Blanchard (2005a) ou Hervouet (2011), ou pour les événements temporels Barazzutti et al. (2015).
Notre interface est dotée de trois fonctions principales : la navigation entre les ensembles de motifs, la visualisation de statistiques sur des ensembles de motifs et la manipulation des ensembles de motifs notamment pour leur comparaison que nous allons détailler.

Tout d'abord, nous donnons la possibilité à l'expert de créer des ensembles de motifs à partir de contraintes sur les mesures d'intérêt, ou à partir de plusieurs jeux de résultats d'extraction (avec des supports minimums différents par exemples). Pour cela, nous avons doté l'interface d'une algèbre des ensembles qui nous permet de comparer des ensembles de motifs. L'interface permet les opérations suivantes : union, intersection, soustraction, soustraction symétrique. Il est aussi possible de procéder à la caractérisation des ensembles numériques par les indicateurs statistiques suivants : cardinal, minimum, maximum, moyenne, médiane, mode, écart-type. La navigation, la manipulation et la comparaison des ensembles de motifs par l'expert sont alors facilitées.

La navigation entre les ensembles issus de la manipulation des données est permise par des aller-retours entre les différentes représentations. Cela ne correspond pas au terme d'hyperdata (Kopecky et Pedrinaci, 2011) mais serait plus proche du concept d'hyperset au sens d'une navigation inter-ensembles.

Nous présentons dans la suite un canevas des possibilités d'utilisation de l'outil. Dans un premier temps, en post-traitement de la fouille, nous procédons au calcul de la r-confiance pour une série de seuils élémentaires compris entre zéro et un, avec un pas de 0.1. Nous obtenons 10 ensembles de motifs que nous comparons via l'interface de l'outil.

FIG. 1: Représentation d'un ensemble de motifs séquentiels sous forme de tableau

## 3.1 Comment créer des ensembles de motifs ?

Nous considérons ici les motifs associés aux valeurs de leurs mesures d'intérêt pour un seuil de confiance élémentaire de 0.9. Nous obtenons la visualisation présentée sur la figure 1. Afin de générer des ensembles de motifs, nous appliquons diverses contraintes. Pour cela, nous disposons d'une console minimale avec un jeu de commandes. On peut pour un premier exemple rechercher les motifs contenant l'item 05M13T (correspondant à une hospitalisation pour douleur thoracique). Soit M l'ensemble des motifs :

$A = \{M_i | 05M13T \preceq M_i\}$

Nous décrivons cette commande ainsi. L'ensemble A est constitué des motifs du premier fichier chargé tel qu'ils incluent l'item 05M13T.

$B = \{M_i \in A | \textit{r-conf}(M_i) > 0.5\}$

Nous obtenons ici un ensemble B constitué des motifs issus de A et respectant la contrainte de support.

Nous pouvons maintenant appliquer des opérations sur ces deux ensembles. Par exemple si nous désirons obtenir le complémentaire de B dans A. Nous procédons de la façon suivante.

$C = \{A \triangle B\}$

L'ensemble C sera le résultat de la soustraction symétrique de l'ensemble A avec l'ensemble B, soit le complémentaire de B dans A.

## 3.2 Comment visualiser graphiquement les ensembles de motifs ?

Nous proposons ici deux représentations possibles des ensembles de motifs accessibles via notre interface figure 2 et figure 3.

La figure 2 est l'histogramme de l'ensemble des motifs séquentiels issus de la fouille de données. Chaque barre représente un ensemble de motifs pour une valeur de support. Chaque bloc d'une barre représente les motifs ayant une même valeur de r-confiance. On peut faire

apparaître, par survol de la souris d'un bloc de l'histogramme les informations correspondantes à savoir : la valeur du support, la valeur de la r-confiance, le nombre de motifs constituant le bloc, le nombre de motifs constituant la barre. Afin d'améliorer l'ergonomie visuelle de l'histogramme empilé nous utilisons un code couleur de type arc-en- ciel. Le bloc le plus important en terme de population de motifs sera codé par la couleur rouge ; les blocs les moins importants seront codés en violet.

La figure 3 représente un arbre correspondant à l'agrégation d'un ensemble de motifs séquentiels. Nous procédons comme suit pour réaliser cette agrégation : nous parcourons l'ensemble des motifs séquentiels étudiés, nous extrayons les motifs débutants par un GHM choisi comme racine de l'arbre, nous parcourons les motifs ainsi extraits item par item, nous créons un nœud pour chaque item, nous créons un arc entre deux items successifs, nous itérons cette procédure sur l'ensemble des motifs extraits. Pour obtenir l'ensemble de la figure 3 nous recherchons les motifs contenant le GHM 05K051. Nous obtenons avec ce GHM comme racine un ensemble de neuf motifs. Nous agrégeons ces neuf motifs sur l'arbre de la figure 3. Le dessin de cet arbre a été obtenu avec l'aide de la librairie D3.js (Data-Driven Documents).

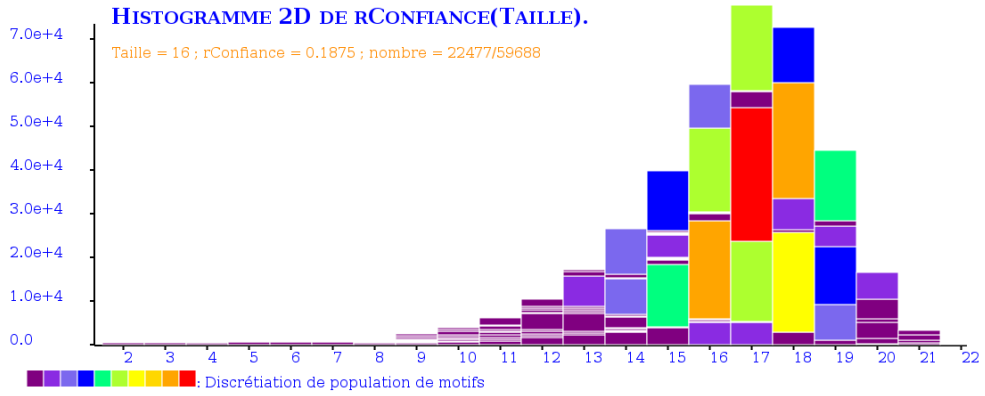



FIG. 2: Représentation d'un ensemble de motifs séquentiels sous forme d'un histogramme empilé

## 3.3 Comment caractériser et comparer statistiquement les ensembles de motifs ?

Les combinaisons de mesures d'intérêt ne sont pas suffisantes pour comparer les grands ensembles de motifs. L'interface développée permet de comparer les ensembles de motifs entre eux à partir d'indicateurs statistiques. Nous procédons ainsi : recherche pour le motif 05K051, codant l'IM. $A = \{M_i | 05K051 \preceq M_i\}, B = \{M_i \in A | \textit{r-conf}(M_i) > 0.2\}$, puis affichage des caractéristiques de A dans la figure 4.

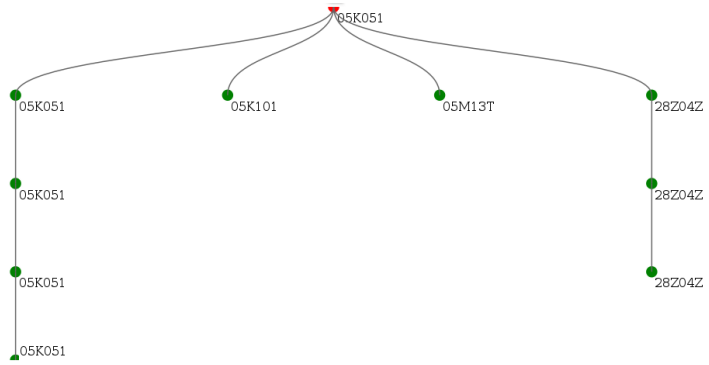



FIG. 3: Représentation d'un ensemble de motifs séquentiels sous forme d'un arbre

**Tableau du Résumé**

| Estimateur | Support | Taille | rConfiance |
|---|---|---|---|
| Cardinal | 9 | 9 | 9 |
| Maximun | 12 | 6 | 0.666667 |
| Minimum | 10 | 3 | 0.4 |
| Moyenne | 10.44 | 4.56 | 0.5 |
| Médiane | 10 | 5 | 0.5 |
| Mode | 10 | 5 | 0.5 |
| Ecart type | 0.68 | 0.83 | 0.09 |

FIG. 4: Caractérisation statistique d'un ensemble de motifs séquentiels

# 4 Application à l'identification de parcours hospitaliers

## 4.1 Données

La collecte des données hospitalières dans le cadre du PMSI génère sur le plan national des bases de données de l'ordre de 25 millions d'enregistrements (séjours) par an [1]. Ces données, recueillies à des fins médico-économiques, peuvent *a posteriori* servir à des fins d'analyse et de recherche, pour examiner des questions médicales et épidémiologiques (Quantin et al., 2014; Bocquier et al., 2011). Dans ces bases, grâce au numéro anonyme de patient, il est possible de reconstituer le parcours hospitalier d'un patient.

Nous nous sommes intéressés aux patients atteints d'un IM, au cours de la période 2009-2013. Pour cela, nous requêtons les bases PMSI nationales 2009-2013, en retenant tous les

1. Guide méthodologique de production des résumés de séjour du PMSI en médecine, chirurgie et obstétrique (fascicule spécial 2004/2 bis du Bulletin officiel)
http://www.atih.sante.fr/textes-officiels-du-pmsi-en-mco

patients ayant un séjour, avec un code acte [2] de cardiologie interventionnelle au cours des cinq années d'observation, soit 490 558 patients (75,7% d'hommes et 24,2% de femmes) [3].

Le taux hospitalier de décès sur la période étudiée est peu élevé (6%). Ce qui peut signifier que cette pathologie est bien prise en charge, ou que le patient décède avant son arrivée à l'hôpital. Pour en savoir davantage sur l'évolution de cette pathologie, nous avons récupéré tous les parcours de soins des patients concernés. À terme, nous souhaitons construire un modèle prédictif en sélectionnant les parcours les plus représentatifs en fréquence, en taille et en confiance. Nous souhaitons à ce titre évaluer la faisabilité d'un parcours lorsqu'un patient a entamé ce parcours. Cette modélisation nous permettra de simuler également à court terme ainsi qu'à long terme, le devenir des patients et les issues possibles de cette pathologie.

Pour notre étude, nous définissons la notion de parcours à l'aide du GHM [4], code renseignant les caractéristiques du séjour hospitalier. À chaque patient, est associé une série de GHM, de longueur égale au nombre de ses séjours effectués sur cinq ans : la trajectoire du patient.

Pour l'extraction de motifs séquentiels, nous avons procédé de la manière suivante : dans un premier temps nous avons écarté les patients ayant moins de 2 séjours, soit 34,4% de la population, afin d'observer une évolution de cette pathologie cardiovasculaire ou autres pathologies éventuelles associées. Ensuite, nous avons créé des contextes à l'aide de variables supplémentaires, qui sont l'âge, le sexe et le nombre d'hospitalisations sur cinq ans. Après concertation avec l'expert clinicien, nous avons discrétisé l'âge en trois classes : les moins de 45 ans, entre 45 et 65 ans et les plus de 65 ans. De même, nous avons décomposé le nombre d'hospitalisation en deux classes : ceux qui viennent entre deux et soixante fois, et ceux qui viennent plus de soixante fois. Toutes combinaisons faites, nous obtenons douze contextes-minimaux. Enfin, nous avons appliqué l'algorithme de recherche de motifs fréquents (Rabatel et al., 2010) sur nos données avec un seuil de 3,5%. Nous obtenons 554 955 motifs fréquents que nous devons trier afin d'en extraire les trajectoires d'intérêts à partir desquelles nous construirons un modèle prédictif par contexte.

Dans la suite, après avoir présenté la méthode, nous allons la mettre à l'épreuve, en tentant de répondre aux questions suivantes :

— **Q1** : Y a t'il une hospitalisation pour IM consécutivement à une hospitalisation pour douleur thoracique ?
— **Q2** : Qu'est-ce qui est fréquemment associé à l'insuffisance cardiaque ?

## 4.2 Exploration des motifs pour réaliser des découvertes médicales

Dans cette section, nous allons utiliser l'outil décrit précédemment pour répondre à la question **Q1**. Nous procédons de la façon suivante : nous recherchons parmi les motifs fréquents, les ensembles de motifs qui contiennent les hospitalisations pour douleur thoracique (code GHM 05M13T) et IM (05K051), sans se soucier de l'ordre dans un premier temps, et nous calculons leur r-confiance pour chacun d'eux pour un seuil élémentaire de 0.9. Nous obtenons deux motifs avec les résultats résumés dans le tableau 2 ci-dessous :

Nous constatons que pour les deux motifs la r-confiance est nulle. Ainsi, lorsque nous recherchons s'il est fréquent d'être hospitalisé pour IM juste après une hospitalisation pour

2. Article L. 6113-8 du code de la santé publique
3. Base nationale : 44,4% d'hommes et 55,6% de femmes
4. Groupe Homogène de Malades

TAB. 2: Résultat de la recherche des sous-ensembles de motifs contenant 05M13T et 05K051

| séquences | support | taille | r-conf |
|---|---|---|---|
| <(05K051)(05M13T)> | 215 | 2 | 0 |
| <(05M13T)(05K051)> | 286 | 2 | 0 |

douleur thoracique (deuxième motif séquentiel du tableau), nous constatons que ce n'est pas le cas. Il est plus probable d'être réhospitalisé pour un (ou plusieurs) autre(s) motif(s) avant de revenir pour IM. Ce qui est cohérent avec les connaissances médicales. En effet, lorsqu'un patient vient pour une douleur thoracique, si cette douleur n'est pas étiquetée IM c'est une alerte. En revanche, le patient peut présenter d'autres complications liées à l'athérosclérose (artériopathie des membres inférieurs, accident vasculaire, *etc.* ) , à l'hypercholestérolémie, au diabète, à l'hypertension, à une insuffisance cardiaque gauche, *etc.* avant de revenir cette fois pour un IM.

Pour répondre à la question **Q2**, cette fois-ci, nous recherchons dans l'ensemble des motifs fréquents les motifs qui contiennent le GHM 05M06, codant l'insuffisance cardiaque. Nous extrayons 44 motifs avec des combinaisons de séquences de séances d'hémodialyse (28Z04Z) de longueur variable. Un parcours retient notre attention : <( 28Z04Z )( 05M043 )( 28Z04Z )> avec une r-confiance de 0.67. Ce qui signifie que si le patient parcourt 1/3 du motif dans 90% des cas il le terminera. En d'autres termes, lorsque le patient est hospitalisé pour une hémodialyse suivi d'une insuffisance cardiaque, il a 90% de chance d'être à nouveau hospitalisé pour une hémodialyse. Il est important de noter que ce motif a été repéré car sa r-confiance est élevée.

En effet, le rôle de l'hémodialyse est d'enlever au patient l'eau que son organisme n'arrive pas à évacuer naturellement. On donne au patient sujet à ces complications un traitement permettant de palier à cette difficulté et un régime alimentaire à suivre. Cependant, si l'observance du patient au regard de son traitement n'est pas bonne, il peut reprendre du poids, c'est essentiellement de l'eau que le coeur ne peut pas traiter et déclencher une insuffisance cardiaque.

# 5 Conclusion

Dans cette étude, nous avons proposé une nouvelle mesure d'intérêt qui est une extension aux motifs séquentiels de la confiance définie pour les règles d'association. Nous avons également développé une interface permettant d'explorer les motifs en prenant en compte différentes mesures d'intérêt dont la r-confiance que nous avons proposée. Nous avons utilisé cet outil pour faire émerger des connaissances médicales à partir d'une base issue des données du PMSI traitant de l'infarctus du myocarde. Les connaissances obtenues ont été validées et décryptées par un expert médical.

L'environnement développé et son utilisation par des utilisateurs experts ont permis de soulever plusieurs limites. La première concerne les contraintes de navigation entre les types de représentation. Par exemple, il n'est pas possible de naviguer directement entre les différentes représentations d'un ensemble de motif. Une deuxième limite concerne l'affichage des ensembles de motifs sous la forme d'un arbre. Pour l'instant, l'information sur les valeurs

d'une mesure d'intérêt n'est pas observable sur les arcs. Pour finir, nous prévoyons de consulter un expert en cardiologie pour évaluer l'impact de la mesure proposée dans la validation des connaissances extraites et sur l'accompagnement dans la découverte de nouvelles connaissances, réel objectif d'une telle plateforme.

# Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, pp. 207–216.

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, et A. I. Verkamo (1996). Advances in knowledge discovery and data mining. Chapter Fast Discovery of Association Rules, pp. 307–328. Menlo Park, CA, USA: American Association for Artificial Intelligence.

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, Washington, DC, USA, pp. 3–14.

Barazzutti, P.-L., A. Cordier, et B. Fuchs (2015). Transmute: an Interactive Tool for Assisting Knowledge Discovery in Interaction Traces. Research report, Universite Claude Bernard Lyon 1 ; Universite Jean Moulin Lyon 3.

Blanchard, J. (2005a). *An interactive visualization system for mining, assessing, and exploring association rules.* Theses, Université de Nantes.

Blanchard, J. (2005b). Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association.

Bocquier, A., N. Thomas, J. Zitouni, E. Lewandowski, S. Cortaredona, M. Jardin, O. Favier, S. Finkel, F. Champion, A. Bernardy, A. Trugeon, et P. Verger (2011). Evaluation of hospital stays linkage quality to study health spatial variation. a feasibility study in three french regions. *Revue d'épidémiologie et de santé publique 59*(4), 243–249.

Charnois, T., M. Plantevit, C. Rigotti, et B. Crémilleux (2009). Fouille de données séquentielles pour l'extraction d'information dans les textes. *Traitement Automatique des Langues(TAL) 50*.

Das, G., K. Lin, H. Mannila, G. Renganathan, et P. Smyth (1998). Rule discovery from time series. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), New York City, New York, USA, August 27-31, 1998*, pp. 16–22.

Dong Haoyuan, L., A. Laurent, et P. Poncelet (2009). Extraction de comportements inattendus dans le cadre du web usage mining. *Revue Nouvelles Technologies de l'Information (RNTI) 2ème Numéro Spécial : Fouille de Données Complexes (2009)*, 113–132.

Etienne, L. et T. Devogele (2012). Mesures de similarité de trajectoires basées sur l'utilisation de patrons spatio-temporels. *Ingénierie des Systèmes d'Information 17*(1), 11–34.

Grissa, D. (2013). *Behavioral study of interestingness measures of knowledge extraction*. Theses, Université Blaise Pascal - Clermont-Ferrand II ; Université de Tunis-El Manar (Tunisie).

Heas, P. (2005). *Apprentissage bayésien de structures spatio-temporelles (application à la fouille visuelle de séries temporelles d'images de satellites)*. Theses, Ecole nationale supé-

rieure de l'aéronautique et de l'espace, Toulouse.

Hervouet, D. (2011). Visualisation des règles d'association en environnement virtuel 3D interactif. Master's thesis.

Kopecky, J. et C. Pedrinaci (2011). RESTful write-oriented API for hyperdata in custom RDF knowledge bases. pp. 199 – 204. IEEE.

Lenca, P., P. Meyer, B. Vaillant, P. Picouet, et S. Lallich (2003). Evaluation et analyse multicritère des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information (RNTI)*, 220–246.

Pei, J., J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M.-C. Hsu (2001). Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. *2014 IEEE 30th International Conference on Data Engineering 0*, 0215.

Pennerath, F. et A. Napoli (2006). La fouille de graphes dans les bases de données réactionnelles au service de la synthèse en chimie organique. In C. D. Gilbert Ritschard (Ed.), *6èmes Journées Francophones "Extraction et gestion des connaissances" - EGC 2006*, Volume 2/RNTI-E-6, Lille, France, pp. 517–528. Cépaduès-éditions.

Quantin, C., Éric Benzenine, M. Hägi, B. Auverlot, M. Abrahamowicz, J. Cottenet, Évelyne Fournier, C. Binquet, D. Compain, Élisabeth Monnet, A.-M. Bouvier, et A. Danzon (2014). Évaluation du pmsi comme moyen d'identification des cas incidents de cancer colorectal. *Santé Publique 26*(1), 55–63.

Rabatel, J. (2011). *Extraction de motifs contextuels : Enjeux et applications dans les données séquentielles)*. Theses, Université MontpellierII Sciences et Techniques du Languedoc.

Rabatel, J., S. Bringay, et P. Poncelet (2010). Aide à la décision pour la maintenance ferroviaire préventive. In *Extraction et Gestion des Connaissances*, EGC'10, Revue des Nouvelles Technologies de l'Information, pp. 363–368. Cépaduès-Éditions.

Sundaram, V. M., A. thnagavelu, et P. Paneer (2012). Discovering co-location patterns from spatial domain using a delaunay approach. *Procedia Engineering 38*, 2832 – 2845. ICMOC12.

Tan, P.-N., V. Kumar, et J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, New York, NY, USA, pp. 32–41. ACM.

# Summary

Sequential patterns mining consist in identifying frequent sequences of ordered events. To solve the problem of the large number of patterns obtained, we extend the interest measure called confidence, conventionally used to select association rules to sequential patterns. In this paper, after presenting the data, we formally define the notion of confidence applied to the sequential patterns. We will apply this measure to identify hospital trajectories, represented by the sequential patterns in data from the PMSI (Medicalization Programme of Information Systems). We focused on a case study: myocardial infarction (MI), in order to predict the trajectory of patients with MI between 2009 and 2013. The results were submitted to an expert for discussion and validation.

# Analyse multi-échelles de référentiels vectoriels via SOLAP

Marie-Dominique Van Damme*, Sébastien Mustière*

*Université Paris Est, IGN-COGIT, Saint-Mandé, France.
marie-dominique.vandamme@ign.fr, sebastien.mustiere@ign.fr

**Résumé.** La France métropolitaine compte aujourd'hui environ 42 millions de bâtiments ce qui représente une surface de 563 000 hectares couvrant ainsi 1% de la surface du territoire. Différentes études ont prouvé l'intérêt des outils décisionnels d'exploration et de visualisation de type SOLAP, pour explorer de larges volumes de données hétérogènes. Dans ce cadre, nous présentons une approche expérimentale à travers les architectures des entrepôts de données spatiales afin d'analyser et agréger, via des requêtes ad-hoc, les données à grande échelle et vectorielles de l'IGN. Ce travail permet déjà à travers des zooms sur les dimensions spatiales et des sélections temporelles de découvrir différents phénomènes et de surligner des caractéristiques sur les protocoles de saisies.

## 1 Introduction

Force est de constater que de plus en plus de données géographiques sont disponibles, que ce soit en raison des progrès techniques et de la baisse du coût de constitution des données, de la mise en place d'infrastructures nationales et internationales, des politiques de diffusion des données publiques, du développement des données dites ouvertes (OpenData), du développement des projets de construction collaborative (Crowdsourcing), ou des possibilités ouvertes par les nouvelles approches du Web (LinkedData). Parmi ces données, on peut noter que même des données à grande échelle sur des territoires étendus, recensant typiquement toutes les routes ou tous les bâtiments de France, sont disponibles, par exemple via les référentiels nationaux de l'IGN ou via les données d'OpenStreetMap. Ces données sont importantes pour la gestion des politiques publiques ou la production de connaissances en général, par exemple pour contrôler l'évolution du territoire dans un contexte de développement durable. Dans ce contexte se pose la question de l'utilisabilité de ces données, détaillées mais volumineuses, pour réaliser des analyses à divers niveaux d'échelle, de l'étude locale à l'étude de tendances sur le territoire national entier. Cette problématique générale renvoie aux questions suivantes : comment manipuler, modéliser et interroger de tels volumes de données afin de réaliser des analyses à des échelles aussi variées ?

La finalité d'un entrepôt de données (DW pour Data Warehouse) est de supporter l'analyse en ligne (OLAP) qui utilise les techniques multidimensionnelles (agrégation partielle des données suivant différents critères). Les données sont stockées sous forme d'hypercube, où chaque dimension représente les axes d'observation et les cellules contiennent les mesures, c'est-à-dire les indicateurs à observer. Les dimensions sont organisées de façon hiérarchique : chaque niveau de la hiérarchie définit la granularité de l'information. Les informations stockées

correspondent aux valeurs des mesures détaillées pour chacune des dimensions. La navigation permet de visualiser les informations contenues dans le cube et de passer d'un niveau d'agrégat à un autre.

Les solutions des entrepôts de données spatiales résultent de la fusion des technologies des DW et des SIG. Elles s'appuient sur une extension du modèle en étoile pour les données spatiales dans les dimensions (Malinowski et Zimányi (2005)) ou pour introduire de nouvelles mesures agrégatives spatiales (Malinowski et Zimányi (2004)). Les dimensions spatiales sont caractérisées par l'existence dans une hiérarchie d'un niveau dont l'une des propriétés des membres est géométrique. Par défaut la relation qui permet d'agréger est la même que celle des dimensions thématiques c'est la relation topologique d'inclusion. Mais il en existe d'autres : l'intersection, la juxtaposition, etc. SOLAP est définit par Bédard et al. (2006) comme une plateforme visuelle combinant des outils OLAP, des fonctionnalités SIG et des outils de géovisualisation. Des outils cartographiques de restitution ont aussi été proposés (Bimonte (2014)) afin d'étendre les normes de géovisualisation. Dans ce domaine des techniques d'optimisation et d'indexation ont été étudiées (Papadias et al. (2001)). Encore récemment des études démontrent l'intérêt, l'actualité et l'apport de ces outils : étude sur des réseaux hydrographiques (Boulil et al. (2014)), intégration des données spatiales vagues (Siqueira et al. (2014)), etc.

Dans cet article, nous présentons une approche SOLAP mise en œuvre pour étudier les bâtiments du Référentiel à Grande Échelle (RGE) de l'IGN. La partie suivante détaille les données utilisées. La partie 3 définit la modélisation SOLAP choisie et la partie 4 illustre l'analyse qui en est faite.

## 2 Données utilisées

L'IGN est en charge de constituer le référentiel à grande échelle (RGE). Cette base de données décrit de façon précise et homogène le territoire français d'un point de vue géométrique et physique. Dans cette base de données vectorielles, chaque bâtiment est individualisé et représenté par sa géométrie ainsi que d'un ensemble d'attributs, comme sa nature ou sa hauteur. Les données actualisées et exhaustives sont contrôlées afin de respecter les spécifications du produit.

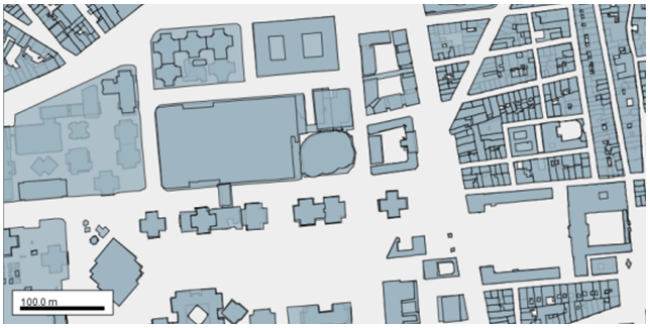



FIG. 1 – *Bâtiments du thème BATI de la BDTOPO*

En particulier, la composante topographique (BDTOPO) contient le thème du bâti dont le degré d'exhaustivité exigé est de 95%. Cette cartographie très détaillée et complète offre donc la possibilité de faire un suivi et un inventaire du territoire d'un point de vue macroscopique mais aussi à différentes échelles. Avant d'être publiées et archivées, les données sont stockées dans une base de données. Cette base contient l'historique des données saisies au fil du temps, aussi bien les bâtiments présents sur le territoire que ceux supprimés : soit parce qu'ils n'existent plus sur le terrain, soit parce qu'ils ont été remplacés (figure 1).

Afin de définir la dimension spatiale qui permettra de ventiler les résultats géographiquement, se pose la question de quel découpage géographique choisir. Les politiques publiques et plus particulièrement celles de la ville s'appuient en général plutôt sur des résultats au niveau de l'infra communale. C'est ce qui a décidé à choisir les Iris (Îlots Regroupés pour l'Information Statistique) définis par l'INSEE [1] comme le découpage le plus fin pour notre étude.

# 3 Définition du cube Bâti

Étant donné qu'il s'agit d'une étude de faisabilité, le travail commence par une première étape de modélisation qui pourra dans une seconde version s'améliorer et s'optimiser, ce qui sera détaillé ensuite.

## 3.1 Modélisation du cube

Cette partie a pour objectif de présenter la démarche mise en œuvre pour modéliser les données de la composante topographique du RGE : la BDTOPO de l'IGN [2]. Nos premières analyses se sont centrées autour des données sur le thème du Bâti. Des travaux antérieurs réalisés au COGIT ont permis d'avoir de nombreuses mesures spatiales sur ces différents objets géographiques (Bard (2004)), et plus particulièrement au niveau du bâtiment, dans un but de comparer la qualité de généralisations.

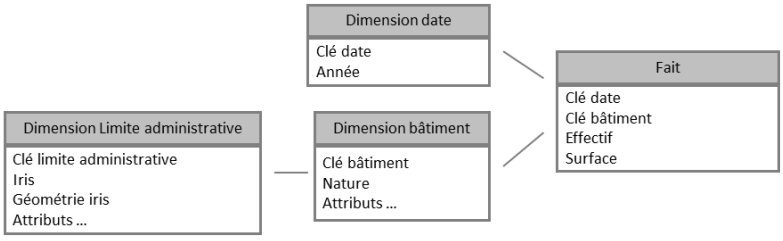



FIG. 2 – *Modèle multidimensionnel "Bâtiment"*

La première étape de la modélisation consiste à définir le niveau de granularité de la table de faits. Dans notre cas, la donnée la plus granulaire est la présence d'un bâtiment en France métropolitaine pour une année. Un premier modèle multidimensionnel (figure 2) est ainsi défini

1. http://professionnels.ign.fr/sites/default/files/Supplements_Gratuits.pdf
2. http://professionnels.ign.fr/sites/default/files/DC_BDTOPO-2-1.pdf

suivant 2 dimensions : la dimension "bâtiment" et la dimension "année" et les mesures qui apparaîtront dans la table de faits sont : effectif des bâtiments et la taille totale (surface en $m^2$).

1. Dimension « Bâtiment »

   Cette dimension thématique contient, en plus de la géométrie du bâtiment, des données qualitatives : la nature (qui permet de différencier les types de bâtiments suivant leur architecture, comme par exemple « Indifférencié », « Château », « Industriel, agricole ou commercial », etc.), la source de la géométrie (qui permet de connaître la source d'information qui a directement servi à la localisation des objets géométriques, comme par exemple "restitution photogrammétrique", "intégration des données du cadastre") et des données quantitatives, comme par exemple la hauteur, l'altitude.

   A cela peuvent s'ajouter des caractéristiques concernant la forme d'un bâtiment qui peuvent se déduire de la géométrie des objets par analyse spatiale :
   - Forme : élongation, concavité (rapport entre la surface du bâtiment et la surface de son enveloppe convexe), nombre de murs, granularité (longueur du plus petit mur)
   - Orientation générale (à $\pi$ près) : l'orientation du plus petit rectangle englobant.
   - Orientation des murs (à $\pi/2$ près) : mesure utilisée pour comparer l'orientation des murs entre différents bâtiments. Son calcul est détaillé dans Duchêne et al. (2003).

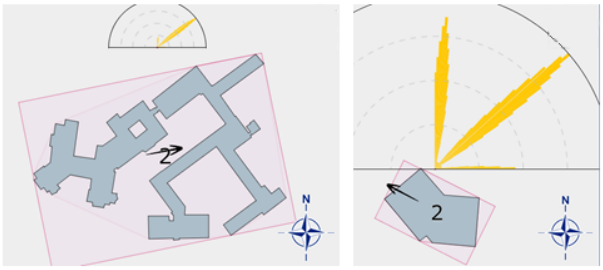



FIG. 3 – *Orientation générale (flèche noire) et orientation des murs (histogramme circulaire), exemple sur deux bâtiments*

   La figure 3 montre les deux orientations de deux bâtiments : la flèche noire indique l'orientation générale et l'histogramme circulaire les différentes orientations des murs.

2. Dimension « Temps »

   La dimension temporelle permet de suivre l'évolution de la base depuis 2006. La granularité de cette dimension est l'année. Cela n'aurait en effet pas de sens de descendre à un niveau inférieur, les mises à jour étant assez discontinues dans le temps. Les mesures ne sont pas agrégatives sur cette dimension (semi-additives).

3. Dimension « Limite administrative »

   Un bâtiment, s'il est présent pour une année, est contenu dans un Iris. L'Iris est un attribut de la dimension "Bâtiment". Mais si l'on devait construire d'autres cubes OLAP avec des données de la BDTOPO (par exemple celui des routes), nous aurions de même un attribut "Iris". La normalisation des dimensions permet de créer une dimension propre "Limite administrative" dont la clé est l'identifiant d'un Iris. Cette dimension peut alors se partager avec d'autres cubes (on parle de flocons de neige).

Cette dimension est importante car c'est elle qui permet de calculer les distributions spatiales. Elle est composée d'une seule hiérarchie de plusieurs niveaux : iris regroupés en communes, qui à leur tour sont regroupées en cantons, qui sont regroupés en départements eux-mêmes regroupés en régions administratives. Le dernier niveau, appelé "All" recouvre la France métropolitaine.

## 3.2 Optimisation du modèle

En adoptant ce premier modèle, la table "fait" centrale devrait contenir un peu plus de 317 millions de lignes (une ligne par couple année-bâtiment) et la table bâtiment 60 millions de lignes (une par bâtiment) pour un espace disque de 60 Go (attributs compris). Nous avons constaté que l'analyse avec ce modèle prend trop de temps.

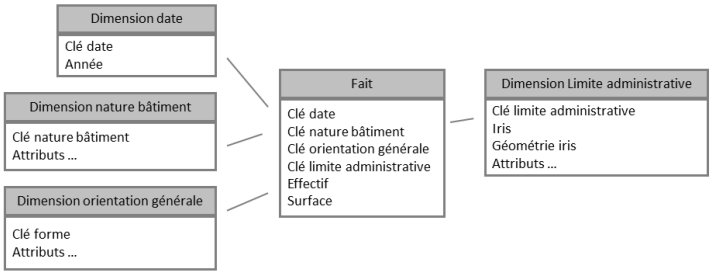



FIG. 4 – *Modèle multidimensionnel "Groupe Bâtiment"*

Une seconde solution consiste à modifier la granularité du cube. L'objectif étant d'agréger à des échelles plus petites, il est préférable de prendre un niveau de granularité moins fin. La dimension "bâtiment" a donc été remplacée par les dimensions "nature", "source de la géométrie", "orientation générale", etc. (cf. figure 4).

L'essentiel des analyses sont cartographiques, l'attribut "Iris" est donc très souvent sollicité. Une technique d'optimisation consiste à rattacher la dimension "Limite administrative" à la table de "faits".

La table de "faits" contient finalement environ 75 millions de lignes (une ligne par groupe de bâtiments partageant les mêmes caractéristiques (date,nature,orientation générale, etc.)).

## 3.3 Mise en œuvre du cube

L'architecture choisie pour implémenter notre solution est calquée sur une architecture type ROLAP combinant différents outils opensource. L'approche ROLAP siginifie que le cube OLAP s'appuie sur un SGBD relationnel classique muni d'un moteur supplémentaire OLAP qui va fournir la vision multidimensionnelle de l'entrepôt ainsi que les calculs dérivés et agrégés.

- L'outil ETC (extraction-tranformation-chargement, ETL en anglais), d'alimentation de l'entrepôt de données, a été implémenté avec la librairie GeOxygene [3] (Bucher et al.

3. `http://oxygene-project.sourceforge.net/`

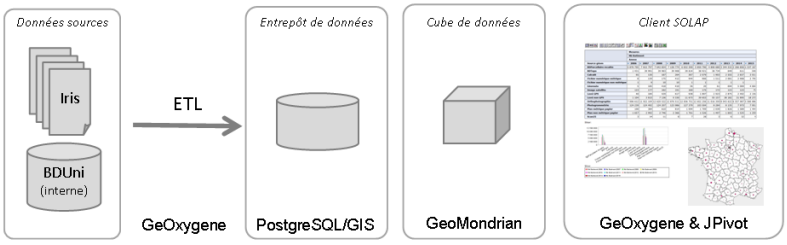



FIG. 5 – *Architecture*

(2009)). En effet celle-ci intègre tous les algorithmes géométriques nécessaires au calcul des caractéristiques du bâtiment. L'opération de jointure entre le centre des bâtiments et la couche des limites des Iris est réalisée dans cette phase. Quelques pertes ont été enregistrées à la fin du processus (géométrie non valide, bâtiment hors iris, etc).

- Un SGBD relationnel pour stocker les données : PostGIS. La table "fait" contient au final environ 75 millions de lignes.
- Un serveur SOLAP : GeoMondrian [4] basé sur Mondrian [5], le moteur OLAP opensource de Pentaho. Ce serveur SOLAP permet de stocker dans le cube de données des géométries (polygone, etc.) comme des propriétés de membres. Il permet aussi d'interroger le cube de données en intégrant des opérateurs spatiaux via le langage MDX (langage de requête pour interroger une structure OLAP).
- Le client SOLAP choisi est le client cartographique de GeOxygene, les résultats étant présentés essentiellement de manière cartographique. Les requêtes MDX sont envoyées au serveur OLAP en respectant le standard XMLA. Les outils JPivot et le requeteur MDX intégrés à la suite GeoMondrian ont permis de vérifier les résultats obtenus.

Voici un exemple d'une requête MDX permettant de calculer le nombre de bâtiments pour deux classes d'orientation ([0, $\pi$/18[ et [$\pi$/18, 2$\pi$/18[) pour le département du Rhône.

```
WITH SET [modalite] AS {[Orientation Mur].[total].[[0, pi/18[],
                [Orientation Mur].[total].[[pi/18, 2pi/18[]}
     MEMBER [Measures].[geom] AS
                '[Limite ilots Regroupes INSEE].CurrentMember.Properties("geom")'
    MEMBER [Measures].[Ventil_Ordre] AS '
                IIF([Orientation Mur].CurrentMember.Level.Ordinal = 0, 1,
            Cast([Orientation Mur].CurrentMember.Properties("ordre") AS INTEGER))'
    MEMBER [Measures].[Portion] AS  [Measures].[Nb Batiment]/SUM([modalite])*100
SELECT {[Measures].[Nb Batiment], [Measures].[Portion], [Measures].[geom_pt]} ON COLUMNS,
   Crossjoin({[Limite ilots Regroupes INSEE.dim_iris].[France entiere].[RHONE-ALPES].[RHONE]},
   ORDER( {[Orientation Mur].[total].[[0, pi/18[], [Orientation Mur].[total].[[pi/18, 2pi/18[]},
              [Measures].[Ventil_Ordre], "DESC") ) ON ROWS
FROM [Cube Batiment Forme]
```

4. http://www.spatialytics.org/fr/projets/geomondrian/
5. mondrian.pentaho.com

# 4 Analyse SOLAP

OLAP désigne un outil graphique qui permet d'analyser et de visualiser les données de manière interactive et rapide en interrogeant de manière itérative le cube par des forages dynamiques et permettant des changements d'axe à la demande. Nous présentons dans cette section deux premiers types de résultats, le premier sur l'axe spatial et le second sur l'axe temporel.

## 4.1 Analyse générale de la forme des bâtiments

Sur l'interface cartographique, la première vue que l'utilisateur obtient est une information avec la plus grande granularité, qui correspond à "France entière" donc. La figure 6-a donne la distribution des effectifs de bâtiments par classe d'orientation générale. La répartition est très homogène à ce niveau d'agrégation. L'intérêt d'un tel outil permet d'utiliser les opérations de forage spatial qui ventilent l'information à un niveau inférieur. Avec un premier "spatial drill-down" les résultats au niveau de la région administrative (figure 6-b) montrent déjà quelques tendances : l'orientation générale des bâtiments est très caractéristique et symétrique à 90° près pour les régions longeant l'Atlantique ou la Méditerranée ainsi que pour la région Rhône-Alpes. Cela se confirme en forant au niveau département (figure 6-c).

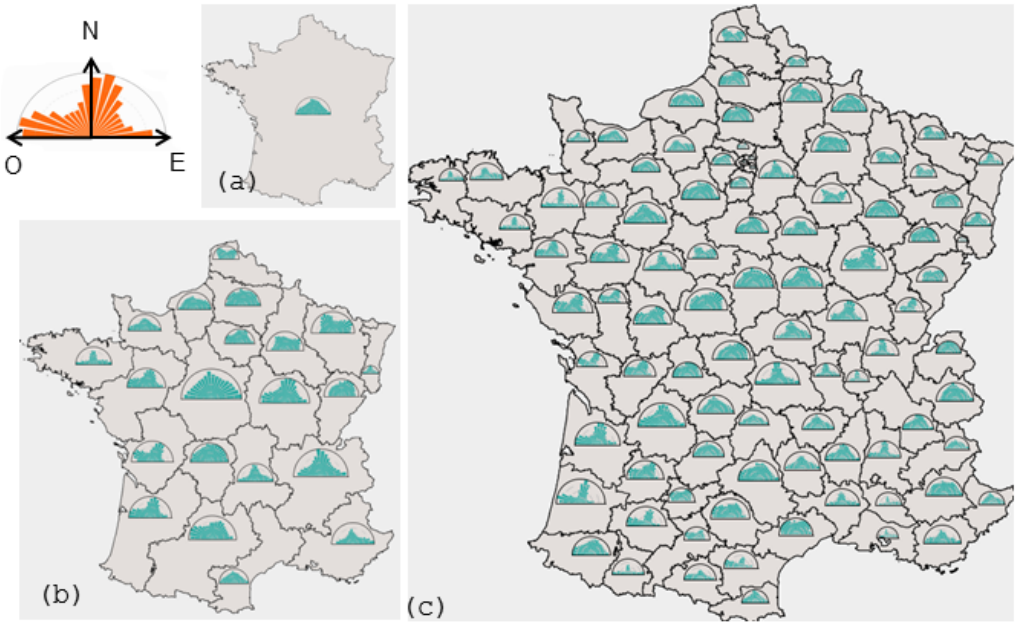



FIG. 6 – *Spatial Drill-Down sur l'orientation générale des bâtiments : (a) France entière (b) Région administrative (c) Département*

L'analyse SOLAP permet également de faire des coupes dans l'hypercube, opération de slicing, comme par exemple en exécutant la même requête que précédemment mais uniquement

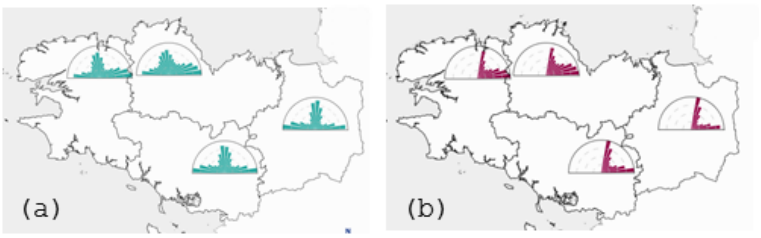



FIG. 7 – *(a) Orientation générale des bâtiments (b) : Orientation principale des murs des bâtiments*

pour les bâtiments situés à +1000m d'altitude. Mais le résultat cartographique ne démontre pas cette fois de tendances. C'est le principe de l'analyse SOLAP. Le système permet d'explorer des hypothèses, fin de les infirmer ou non.

En opérant une opération de pivot, c'est-à-dire en changeant l'axe d'observation "orientation générale" par "orientation principale des murs" sur la région Bretagne (figure 7), à l'échelle départementale la différence semble minimale. Cela peut s'expliquer par le fait que majoritairement les bâtiments sont rectangulaires. Cette dernière hypothèse pourrait être confirmée en rajoutant comme caractéristique "nombre de murs" à la dimension "Bâtiment".

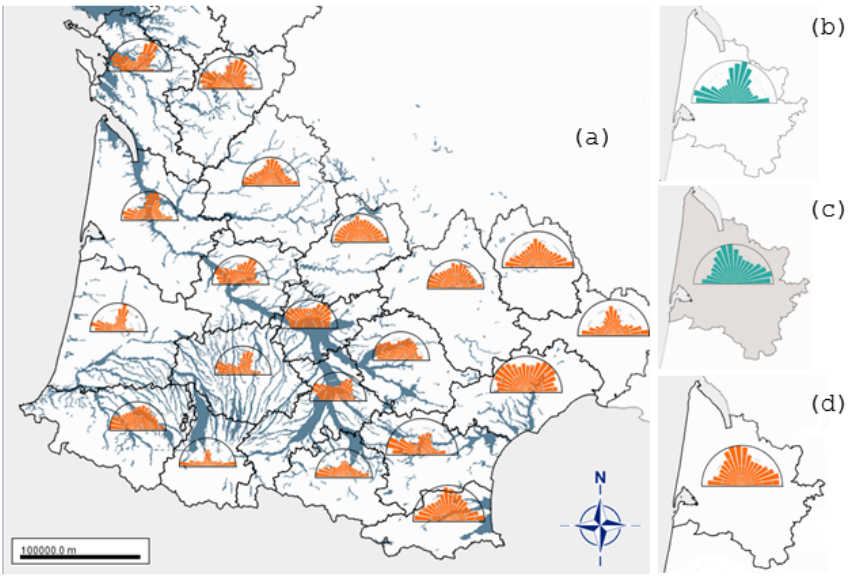


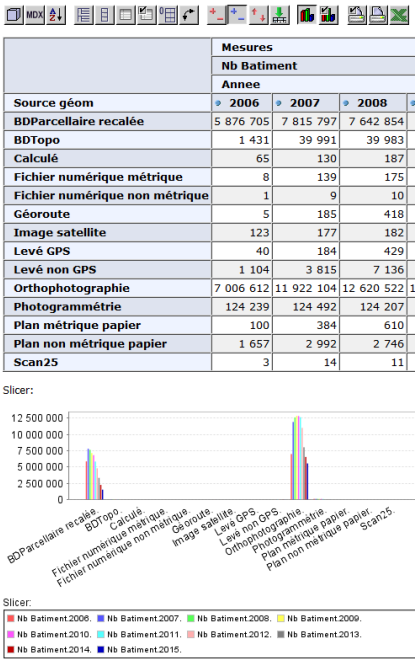

FIG. 8 – *(a) co-visualisation de l'orientation des bâtiments avec en fond les vallées du bassin Adour-Garonne (b) orientation générale des bâtiments en Gironde (c) orientation des routes en Gironde, (d) Orientation de la route la plus proche des bâtiments en Gironde*

Les résultats de la figure 8-a résultent de l'opération dicing, c'est-à-dire d'une sélection sur la dimension géographique des départements des régions Aquitaine, Poitou-Charentes, Midi-Pyrénées et Languedoc-Roussillon depuis la troisième requête effectuée (orientation générale des bâtiments sur France entière par département). En fond cartographique la couche des vallées du bassin Adour-Garonne permet de covisualiser les résultats SOLAP et peut-être d'y voir d'éventuelles relations. Cette approche reste cependant limitée.

Une autre solution visuelle consisterait à mettre côte à côte des résultats similaires afin de les comparer. Par exemple la figure 8 b-c-d confronte l'orientation générale des bâtiments (b), l'orientation des routes de toutes natures (c) et l'orientation de la route la plus proche des bâtiments (d). Mais l'interdépendance de ces variables reste encore trop limitée et doit être analysée plus en profondeur.

## 4.2 Analyse avec la dimension Temps

Le tableau de la figure 9 montre l'évolution des effectifs des bâtiments dans le temps pour la région Aquitaine. Le bond en 2007 correspond à l'intégration des bâtiments du cadastre, donnant une description plus fine des bâtiments, et donc un découpage en plus de bâtiments individualisés. Cette intégration se faisant au fur et à mesure, sur quelques années, l'effectif compte aussi bien des bâtiments issus de la BDTOPO que des bâtiments issus du cadastre.



| | Mesures | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Nb Batiment | | | | | | | | | |
| | Annee | | | | | | | | | |
| Source géom | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
| BDParcellaire recalée | 5 876 705 | 7 815 797 | 7 642 854 | 7 158 774 | 6 832 590 | 5 850 796 | 4 808 680 | 3 344 532 | 2 266 856 | 1 537 107 |
| BDTopo | 1 431 | 39 991 | 39 983 | 39 908 | 39 814 | 38 921 | 38 734 | 649 | 611 | 540 |
| Calculé | 65 | 130 | 187 | 264 | 367 | 2 079 | 1 903 | 2 555 | 2 837 | 3 011 |
| Fichier numérique métrique | 8 | 139 | 175 | 412 | 944 | 906 | 1 511 | 2 081 | 2 908 | 2 795 |
| Fichier numérique non métrique | 1 | 9 | 10 | 10 | 1 | 1 | 1 | 1 | 1 | |
| Géoroute | 5 | 185 | 418 | 410 | 36 | 20 | 61 | 694 | 5 069 | 4 664 |
| Image satellite | 123 | 177 | 182 | 191 | 169 | 175 | 172 | 123 | 123 | 75 |
| Levé GPS | 40 | 184 | 429 | 617 | 938 | 1 897 | 2 923 | 2 875 | 2 492 | 2 156 |
| Levé non GPS | 1 104 | 3 815 | 7 136 | 9 330 | 11 673 | 39 653 | 55 157 | 38 182 | 31 895 | 18 272 |
| Orthophotographie | 7 006 612 | 11 922 104 | 12 620 522 | 12 876 511 | 12 836 751 | 12 652 256 | 11 016 359 | 8 043 622 | 6 557 987 | 5 560 886 |
| Photogrammétrie | 124 239 | 124 492 | 124 207 | 123 986 | 127 379 | 103 504 | 8 284 | 8 135 | 7 973 | 7 951 |
| Plan métrique papier | 100 | 384 | 610 | 814 | 1 094 | 1 709 | 2 039 | 1 826 | 1 589 | 1 444 |
| Plan non métrique papier | 1 657 | 2 992 | 2 746 | 2 366 | 1 761 | 1 526 | 1 497 | 1 603 | 1 525 | 1 259 |
| Scan25 | 3 | 14 | 11 | 8 | 7 | 28 | 3 | 6 | 32 | 9 |



FIG. 9 – *Nombre de bâtiment par année et par source d'information.*

# 5 Discussion et perspectives

Ce travail, encore en cours, nous permet d'avoir des premiers retours d'expérience.

Le premier retour d'expérience nous montre que manipuler de tels volumes de données reste long mais réalisable. Dans nos expériences, l'alimentation des différents cubes envisagés pour traiter la France entière a nécessité en temps quelques jours sur un ordinateur individuel. Ce n'est néanmoins pas rédhibitoire car cette étape n'est faite qu'une seule fois en amont. Concernant l'analyse, le temps de réponse des requêtes prend en moyenne plusieurs dizaines de secondes. Ceci est encore trop long pour une analyse fluide et interactive des données, et cela doit être optimisé en priorité. Les pistes à étudier à ce sujet sont les indexations, des pré calculs de certaines requêtes, ou une répartition des données en sous-cubes.

En termes de modélisation, on identifie deux sujets principaux : le choix de la granularité spatiale et la modélisation du temps.

Au niveau de la granularité, une première possibilité est d'individualiser chaque bâtiment dans la table de faits, mais cela conduit à des gros volumes de données (surtout si les bâtiments sont dupliqués pour chaque année d'existence), ce que confirment nos premières expérimentations, où nous avons atteint les limites des outils utilisés pour constituer la table de bâtiments. Pour aller plus loin dans cette direction, les questions d'optimisation des outils ROLAP deviennent cruciales (Papadias et al. (2001)). Une autre possibilité est de faire des premières agrégations dès la constitution du cube, où les faits atomiques sont les groupes de bâtiments dans un Iris avec les mêmes caractéristiques (nature, orientation, etc.). Nos expériences montrent que c'est plus réalisable avec des temps de requêtes encore longs mais acceptables. Cependant, cela pose le problème de la compréhensibilité du modèle, et de sa stabilité car tout ajout d'une nouvelle dimension oblige à répartir les bâtiments et donc à redéfinir la granularité et le chargement de la table de faits.

Au niveau du temps, plusieurs choix sont aussi possibles : les bâtiments qui perdurent sur plusieurs années peuvent être dupliqués pour chaque année de leur existence. C'est cette approche que nous avons suivi. Mais cela a pour inconvénient de faire exploser la taille de la table de faits et rend plus difficile l'analyse des évolutions individuelles. Une autre solution est de considérer les dates de création et destruction comme deux attributs de la dimension bâtiments. Ce dernier point reste à creuser.

Un autre retour d'expérience concerne l'analyse des corrélations entre deux phénomènes géographiques, comme par exemple dans notre cas l'orientation des bâtiments vis-à-vis de celle des routes ou des rivières. Plusieurs approches sont possibles. On peut afficher un fond de données géographiques comme fond de carte (ex : les rivières) pour comprendre un phénomène. Une autre solution est de constituer deux cubes : un pour l'orientation des routes, un pour les rivières, et de co-visualiser ces cubes. Une troisième solution est de considérer de nouveaux attributs définissant le contexte spatial des bâtiments : par exemple l'orientation de la route ou de la rivière la plus proche. Les forces et faiblesses de ces différentes approches méritent d'être explorées.

Une autre question est la gestion de l'évolution des partitions territoriales : non seulement un bâtiment évolue dans le temps, mais il peut aussi changer d'IRIS au cours du temps. Comment dans ces conditions modéliser l'évolution des découpages spatiaux dans le temps ?

Une interrogation qui reste ouverte également est la transposition de ce genre d'approche à des données issues de collectes collaboratives, qui sont riches mais parfois hétérogènes, et pour lesquelles il est difficile de comparer les quantités calculées. Pour ces données, une gestion des incertitudes et des stratégies d'échantillonnage sont surement à mettre en place.

# Références

Bard, S. (2004). *Méthode d'évaluation de la qualité de données généralisées - applications aux données urbaines*. Thèse de doctorat, spécialité informatique, Université de Paris VI.

Bimonte, S. (2014). A generic geovisualization model for spatial olap and its implementation in a standards-based architecture. *Ingénierie des Systèmes d'Information 19*(5), 97–118.

Boulil, K., F. L. Ber, S. Bimonte, C. Grac, et F. Cernesson (2014). Multidimensional modeling and analysis of large and complex watercourse data : an olap-based solution. *Ecological Informatics 24*, 90 – 106.

Bucher, B., M. Brasebin, E. Buard, E. Grosso, et S. Mustière (2009). Geoxygene : built on top of the expertness of the french nma to host and share advanced gi science research results.

Bédard, Y., S. Rivest, et M. josée Proulx (2006). Spatial on-line analytical processing (solap) : Concepts, architectures, and solutions from a geomatics engineering perspective. In *Data Warehouses and OLAP : Concepts, Architecture, and*, pp. 298319. Press.

Duchêne, C., S. Bard, X. Barillot, A. Ruas, J. Trévisan, et F. Holzapfel (2003). Quantitative and qualitative description of building orientation. In *6th ICA Workshop on Generalisation and Multiple Representation, 28-30 April, Paris (France)*.

Malinowski, E. et E. Zimányi (2004). Representing spatiality in a conceptual multidimensional model. In *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, GIS '04, New York, NY, USA, pp. 12–22. ACM.

Malinowski, E. et E. Zimányi (2005). Spatial hierarchies and topological relationships in the spatial multidimer model. In M. Jackson, D. Nelson, et S. Stirk (Eds.), *BNCOD*, Volume 3567 of *Lecture Notes in Computer Science*, pp. 17–28. Springer.

Papadias, D., P. Kalnis, J. Zhang, et Y. Tao (2001). Efficient olap operations in spatial data warehouses. In *Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases*, SSTD '01, London, UK, UK, pp. 443–459. Springer-Verlag. English

Siqueira, T. L. L., C. D. d. A. Ciferri, V. C. Times, et R. R. Ciferri (2014). Modeling vague spatial data warehouses using the vscube conceptual model. *GeoInformatica 18*(2), 313–356.

# Index

Azé, Jérôme, 34

Bringay, Sandra, 34

Danielle, Ziebelin, 1

Fotsoh Tawofaing,Armel, 21

Guez, Alain, 9

Kergosien, Eric, 2

Le Parc-Lacayrelle, Annig, 21

Mercadier,Yves, 34

Moal,Tanguy , 21
Mustière, Sébastien , 47

Pinaire, Jessica, 34

Roche, Mathieu, 2
Rousseaux, Francis, 9

Teisseire, Maguelonne, 2, 34

Van Damme, Marie-Dominique , 47

Zenasni, Sarah, 2