

Interconnexion et visualisation de ressources géoréférencées du Web de données à l'aide d'un référentiel topographique de support

Abdelfettah Feliachi

▶ **To cite this version:**

Abdelfettah Feliachi. Interconnexion et visualisation de ressources géoréférencées du Web de données à l'aide d'un référentiel topographique de support. Géographie. Université Paris-Est, 2017. Français. NNT : 2017PESC1179 . tel-01787128

HAL Id: tel-01787128
https://theses.hal.science/tel-01787128

Submitted on 7 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Présentée pour obtenir le grade de docteur de l'Université Paris-Est

Spécialité : Sciences et Technologies de l'Information Géographique

ABDELFETTAH FELIACHI

Interconnexion et visualisation de ressources géoréférencées du Web de données à l'aide d'un référentiel topographique de support

Soutenu le 27 octobre 2017

Jury:

Oscar Corcho, Professeur, Universidad Politécnica de Madrid ..... Rapporteur

Nathalie Pernelle, Maître de conférences, HDR, LRI-Université Paris Sud ..... Rapporteur

Jérôme Euzenat, Directeur de recherche, INRIA et Uni. de Grenoble Alpes..... Examinateur

Fabian Suchanek, Professeur, Télécom ParisTech..... Examinateur

Thomas Devogele, Professeur, Université François Rabelais de Tours ..... Examinateur

Bénédicte Bucher, Ingénieur, HDR, COGIT-IGN,..... Directrice de thèse

Nathalie Abadie, Ingénieur, Docteur, COGIT-IGN, ... Encadrante

Fayçal Hamdi, Maître de Conférences, CEDRIC-CNAM,..... Encadrant

Université Paris-Est - École Doctorale MSTIC

Institut National de l'Information Géographique et Forestière

Équipe de Recherche COGIT (Cartographie et Géomatique)

Cette thèse a été réalisée à l'équipe COGIT du Laboratoire LaSTIG de l'Institut National de l'Information Géographique et Forestière, sous la direction de Bénédicte Bucher et l'encadrement de Nathalie Abadie et Fayçal Hamdi.

Institut National de l'Information Géographique et Forestière
Service de la Recherche, Laboratoire LaSTIG
73 Avenue de Paris
94165 Saint-Mandé Cedex

Tél. : 01 43 98 80 00

RÉSUMÉ

Plusieurs ressources publiées sur le Web de données sont dotées de références spatiales qui décrivent leur localisation géographique. Ces références spatiales sont un moyen favori pour interconnecter et visualiser les ressources sur le Web de données. Cependant, les hétérogénéités des niveaux de détail et de modélisations géométriques entre les sources de données constituent un défi majeur pour l'utilisation de la comparaison des références spatiales comme critère pour l'interconnexion des ressources. Ce défi est amplifié par la nature ouverte et collaborative des sources de données du Web qui engendre des hétérogénéités géométriques internes aux sources de données. En outre, les applications de visualisation cartographique des ressources géoréférencées du Web de données ne fournissent pas une visualisation lisible à toutes les échelles.

Dans cette thèse, nous proposons un vocabulaire pour formaliser les connaissances sur les caractéristiques de chaque géométrie dans un jeu de données. Nous proposons également une approche semi-automatique basée sur un référentiel topographique pour acquérir ces connaissances. Nous proposons de mettre en œuvre ces connaissances dans une approche d'adaptation dynamique du paramétrage de la comparaison des géométries dans un processus d'interconnexion. Nous proposons une approche complémentaire s'appuyant sur un référentiel topographique pour la détection des liens de cardinalité n:m. Nous proposons finalement des applications qui s'appuient sur des données topographiques de référence et leurs liens avec les ressources géoréférencées du Web pour offrir une visualisation cartographique multiéchelle lisible et conviviale.

Mots-clés : Web de données, données liées, références spatiales, hétérogénéités, interconnexion de données, visualisation cartographique de données.

ABSTRACT

Interlinking and visualizing georeferenced resources of the Web of data with geographic reference data

Many resources published on the Web of data are related to spatial references that describe their location. These spatial references are a valuable asset for interlinking and visualizing data over the Web. However, these spatial references may be presented with different levels of detail and different geometric modelling from one data source to another. These differences are a major challenge for using geometries comparison as a criterion for interlinking georeferenced resources. This challenge is even amplified more due to the open and often volunteered nature of the data that causes geometric heterogeneities between the resources of a same data source. Furthermore, Web mapping applications of georeferenced data are limited when it comes to visualize data at different scales.

In this PhD thesis, we propose a vocabulary for formalizing the knowledge about the characteristics of every single geometry in a dataset. We propose a semi-automatic approach for acquiring this knowledge by using geographic reference data. Then, we propose to use this knowledge in approach for adapting dynamically the setting of the comparison of each pair of geometries during an interlinking process. We propose an additional interlinking approach based on geographic reference data for detecting n:m links between data sources. Finally, we propose Web mapping applications for georeferenced resources that remain readable at different map scales.

Keywords: Web of data, linked data, spatial references, heterogeneities, data interlinking, data visualisation in maps.

REMERCIEMENT

Je remercie l'IGN de m'avoir donné la chance d'effectuer cette thèse au sein du laboratoire LASTIG. Je remercie tous les agents qui m'ont aidé de près ou de loin et qui ont fait de ma de thèse une expérience passionnante.

Je remercie tout spécialement tout spécialement les membres de l'équipe COGIT pour l'expérience scientifique et personnelle que je considère l'une des plus enrichissante que je jamais eu.

Je tiens remercier sincèrement les rapporteur, Nathalie Pernelle et Oscar Corcho, d'avoir pris le temps de relire ce manuscrit. Je les remercie pour leurs remarques et leurs questions constructives qui m'ont permis d'élargir les perceptives de ce travail. Je remercie tous les membres du Jury de l'évaluation de ce travail et de la discussion très pertinente qu'ils m'ont accordé.

Je remercie très chaleureusement ma directrice de thèse Bénédicte Bucher de son suivi de mon travail et de m'avoir toujours aidé d'avoir du recul sur mon travail, ce qui m'a aidé à avancer avec plus de certitude. Je la remercie également de m'avoir accordé de l'autonomie quand il le fallait.

Je présente un immense remerciement à mon encadrant Fayçal Hamdi pour tout ce qu'il a fait pour moi au niveau scientifique ainsi que niveau personnelle. Je le remercie d'avoir été à mon aide à tout moment et inconditionnellement. Je le remercie de l'encouragement et de la confiance qu'il m'a toujours accordée. J'ai beaucoup appris de ses cotés et je me considère très chanceux d'avoir travaillé avec lui

Je tiens en dessous de tout de présenter un immense remerciement à mon encadrante Nathalie Abadie grâce sans qui ce n'aurai pas arrivé au bout. Les quatre ans où j'ai eu la chance de travailler avec elle était un vrai privilège. Je suis sincèrement reconnaissant pour sa présence et suivi continus, sa patience, sa rigueur, son humeur et son encouragement continue. Je n'aurai jamais espéré mieux comme encadrement et c'est en grande partie grâce à elle que je suis là où j'en suis aujourd'hui.

Je suis reconnaissant envers mes amis de leur soutien moral et leur patience lors de cette aventure. Je remercie finalement ma famille de son soutien inconditionnel et d'avoir toujours cru en moi. C'est pour vous et grâce à vous que j'ai réussi ce travail. Je ne vous remercierai jamais assez.

TABLES DES MATIÈRES

RÉSUMÉ 4

ABSTRACT 5

REMERCIEMENT 6

TABLES DES MATIÈRES 7

INTRODUCTION 11

PARTIE A LIER ET VISUALISER DES RESSOURCES GÉORÉFÉRENCÉES SUR LE WEB 13

1 CONTEXTE ET OBJECTIFS 15

1.1 Web sémantique et Web de données 16

1.2 La place centrale des ressources géoréférencées sur le Web de données 18

1.3 Vers une représentation standard de données géoréférencées sur le Web de données 21

1.3.1 Des standards pour la représentation de l'information géographique 21

1.3.2 Des vocabulaires pour la représentation des références spatiales directes sur le Web de données 23

1.3.3 Un effort de standardisation commun W3C - OGC 25

1.4 Lier les ressources géoréférencées du Web : enjeux et verrous 27

1.4.1 Lier les ressources grâce à leurs références spatiales : bénéfices et applications 27

1.4.2 Hétérogénéités des références spatiales : des bases de données géographiques au Web de données 30

1.5 Géométries, liens et vocabulaires : visualiser les ressources géoréférencées du Web de données 37

1.6 Objectifs de la thèse 41

2 ÉTAT DE L'ART : APPARIEMENT DE DONNÉES GÉOGRAPHIQUES ET INTERCONNEXION DES DONNÉES SUR LE WEB 42

2.1 Appariement de données géographiques 42

2.1.1 Des mesures de distance pour comparer les objets géographiques 44

2.1.2 Stratégies d'appariement de données géographiques 48

2.2 Interconnexion des Linked Data : concept, approches et outils 54

2.2.1 Étapes d'un processus d'interconnexion 54

2.2.2 Approches et outils d'interconnexion de Linked Data 59

2.3 Conclusion de l'état de l'art 65

PARTIE B PROPOSITIONS 69

3 FORMALISATION ET ACQUISITION DES CONNAISSANCES POUR LA QUALIFICATION DES RÉFÉRENCES SPATIALES DIRECTES SUR LE web DE DONNÉES 71

3.1 Un vocabulaire pour décrire la sémantique des XY 71

3.1.1 Métadonnée sur la précision planimétrique des géométries 73
3.1.2 Métadonnée sur la modélisation géométrique 78
3.1.3 Métadonnée sur le caractère vague de certaines entités géographiques 82
3.1.4 Métadonnée sur la résolution géométrique 82
3.2 Peuplement du vocabulaire de la sémantique des XY 84
3.2.1 Peuplement du vocabulaire dans le cas où les métadonnées sur les géométries sont fournies avec les données 85
3.2.2 Peuplement du vocabulaire en l'absence de métadonnées sur les géométries fournies avec les données 87
3.3 Conclusion 97
4 PROPOSITIONS D'APPROCHES D'INTERCONNEXION ET DE VISUALISATION DES DONNÉES GÉORÉFÉRENCÉES SUR LE web DE DONNÉES 99
4.1 Une approche à base de connaissances pour adapter dynamiquement le paramétrage du processus d'interconnexion de données géoréférencées 99
4.1.1 Formalisation de l'approche 100
4.1.2 Description générale de l'approche 101
4.1.3 Mise en œuvre de l'approche d'adaptation dynamique des paramètres de la comparaison des géométries. 107
4.1.4 Test et validation : interconnexion de monuments de la ville de Paris 113
4.2 Utilisation d'un référentiel topographique comme support pour l'interconnexion de données géoréférencées 124
4.2.1 Description générale de l'approche 124
4.2.2 Mise en œuvre et évaluation de l'approche pour l'interconnexion des monuments parisiens 126
4.3 Applications de Visualisation cartographique de données thématiques géoréférencées .. 130
4.3.1 Exploitation des liens d'ancrage entre données thématiques et données topographiques de support pour la visualisation multi-échelle 131
4.3.2 Vers une application d'exploration cartographique multi-échelle générique des sources de données géoréférencées 135
4.4 Conclusion du chapitre 146
CONCLUSION GÉNÉRALE ET PERSPECTIVES 149
LISTE DES FIGURES 155
LISTE DES TABLEAUX 158
BIBLIOGRAPHIE 159
ANNEXES 169
Annexe A 169

INTRODUCTION

Les ressources publiées sur le Web de données sont souvent accompagnées de références spatiales telles que des noms de lieux, des adresses ou des coordonnées géographiques. Comme toute autre propriété associée aux ressources du Web de données, ces références spatiales peuvent être mises à profit pour comparer des ressources dans un processus d'interconnexion. Les approches actuelles de détection de relations de correspondance entre données à références spatiales reposent principalement sur l'hypothèse selon laquelle les ressources décrites par des références spatiales géographiquement proches sont plus susceptibles de représenter une même entité géographique du monde réel. Ainsi, les ressources sont comparées à l'aide de mesures de distance calculées entre leurs références spatiales. Le choix de la mesure à utiliser et du seuil de distance au-delà duquel deux ressources ne sont plus considérées comme potentiellement équivalentes nécessite de disposer de connaissances sur la précision de localisation des références spatiales, sur l'élément caractéristique de la forme des entités géographiques représenté par ces références spatiales et sur le caractère plus ou moins vague des entités géographiques représentées. Ce choix est le plus souvent réalisé par un expert en mise en correspondance de données. Contrairement aux jeux de données géographiques issus de producteurs de données traditionnels comme les agences cartographiques nationales, les jeux de données publiés sur le Web proviennent pour la plupart de sources de données participatives ou sont le fruit de l'agrégation de plusieurs jeux de données. Les références spatiales qu'ils renferment ne sont donc pas toutes nécessairement produites de la même façon: leurs précisions de localisation, l'élément caractéristique de la forme de l'entité géographique prise comme repère de saisie ou encore le type d'entité géographique représentée peuvent varier d'une référence spatiale à l'autre. Dans ce cas, définir et appliquer un même processus d'interconnexion, doté du même paramétrage pour l'ensemble des ressources traitées, peut alors devenir extrêmement complexe, voire engendrer des liens erronés.

Dans cette thèse, nous proposons de formaliser ces connaissances sur les caractéristiques des références spatiales, nécessaires à l'identification d'hétérogénéités potentielles. Nous proposons donc un vocabulaire permettant d'associer à toute référence spatiale utilisée pour localiser une ressource, des métadonnées sur sa précision de localisation, l'élément caractéristique de la forme de l'entité géographique prise comme repère sur le terrain pour sa saisie ainsi que le caractère plus ou moins vague de cette entité géographique. Afin de compléter les données à interconnecter avec ces métadonnées, nous introduisons une approche fondée sur un apprentissage par classification supervisée et un référentiel de données géographiques pour acquérir ces métadonnées.

Disposant de connaissances sur les caractéristiques des références spatiales, nous pouvons déduire automatiquement le niveau d'hétérogénéité que deux références spatiales sont susceptibles de présenter et ainsi adapter l'approche de mise en correspondance utilisée. Nous proposons donc deux approches d'interconnexion de ressources géoréférencées mettant en œuvre ces connaissances. La première met en œuvre un raisonnement à base de règles sur ces connaissances pour adapter dynamiquement le paramétrage de la comparaison des références spatiales lors du processus d'interconnexion. La seconde, utilise les connaissances sur les entités géographiques représentées pour choisir un référentiel topographique de support permettant de compenser les hétérogénéités entre références spatiales lors de leur mise en correspondance.

Interconnecter des données du Web avec des données géographiques de référence permet de les associer à des références spatiales de niveaux de détails divers, facilitant ainsi leur mise en œuvre dans des applications de visualisation cartographique multi-échelle. Nous proposons ainsi deux applications de visualisation de ressources du Web avec des données géographiques de référence. La première, met en œuvre des liens entre ressources du Web de données et données géographiques dotées de références spatiales plus détaillées afin de faire porter par ces dernières l'information thématique apportée par les ressources du Web de données et améliorer la lisibilité des cartes produites à différentes échelles. La seconde exploite les liens entre ressources issues de n'importe quelle source du Web de données et un maillage territorial multi-échelle afin de générer de façon semi-automatique des cartes statistiques à partir des propriétés de ces ressources.

Plan du mémoire

Ce mémoire est composé de deux parties principales A et B.

La partie A décrit le contexte général de l'interconnexion et la visualisation des ressources géoréférencées sur le Web de données. Dans cette partie, le chapitre 1 expose le rôle important que jouent les références spatiales pour l'interconnexion et la visualisation des ressources du Web de données et les difficultés rencontrées dans ce contexte. Il termine par préciser les objectifs de la thèse. Le chapitre 2 dans cette partie dresse un état de l'art qui joint les différentes approches de l'état de l'art en matière d'appariement de données géographiques et d'interconnexion des données du Web.

La partie B rassemble les différentes propositions que nous avons mises au point dans le cadre de cette thèse. Dans cette partie, le chapitre 4 décrit notre proposition de formaliser ces connaissances sur les caractéristiques des références spatiales et détaille l'approche mise en œuvre pour leur acquisition. Le chapitre 4 regroupe nos différentes propositions pour l'interconnexion et la visualisation des ressources géoréférencées.

Enfin, nous proposons une dernière section qui présente les conclusions générales et les perspectives de ce travail.

PARTIE A
LIER ET VISUALISER DES RESSOURCES GÉORÉFÉRENCÉES SUR LE WEB DE DONNÉES

Le rôle central du niveau de détail et de la modélisation géométrique des données

Introduction

Cette partie décrit le contexte dans lequel s'inscrit le travail de cette thèse qui se situe entre le domaine du Web sémantique et le domaine des sciences de l'information géographique.

Nous nous intéressons aux enjeux et verrous de l'utilisation des références spatiales directes sur le Web de données pour interconnecter et visualiser les ressources. En particulier, nous nous intéressons à deux aspects de la définition des références spatiales directes, intrinsèquement liés l'un à l'autre : leur niveau de détail et leur modélisation géométrique.

Les différences des niveaux de détail et de modélisation géométriques sont à l'origine de divers types d'hétérogénéités entre bases de données géographiques. Ces hétérogénéités ont été identifiées de longue date dans le domaine de l'appariement de données géographiques (Lemarié et Raynal, 1996 ; Devogele et al., 1998) et nous verrons comment les algorithmes d'appariement de la littérature tentent d'y remédier. Dans le domaine de l'interconnexion des ressources du Web de données, le recours aux références spatiales comme critère de mise en correspondance tend à se généraliser. Les outils d'interconnexion proposant ce critère doivent alors faire face à ces mêmes types d'hétérogénéités géométriques, amplifiées par la nature ouverte et l'origine souvent collaborative des sources de données du Web. Nous verrons donc quelles stratégies sont proposées pour automatiser le paramétrage et l'adaptation dynamique des outils d'interconnexion aux hétérogénéités rencontrées.

Par ailleurs, la notion de niveau de détail, qui est le pendant dans le domaine des bases de données géographiques de celle d'échelle en cartographie, joue un rôle central dans la mise en œuvre de solutions de visualisation cartographique de données. En effet, si les références spatiales associées aux données permettent d'en proposer une visualisation cartographique, celle-ci reste limitée par des contraintes de lisibilité liées au niveau de détail des géométries et à l'échelle d'affichage de la carte.

1 CONTEXTE ET OBJECTIFS

Cette partie a pour objectif de décrire le contexte général de cette thèse : les ressources géoréférencées publiées sur le Web de données, ainsi que les enjeux et les difficultés que représentent leur représentation et leur interconnexion pour la constitution du Web de données. Nous verrons à travers ce chapitre que ce contexte est fortement lié au domaine des bases de données géographiques.

Dans le Web de données, On appelle une « **ressource** » toute entité concrète ou abstraite pouvant être désignée par un identifiant. Ceci inclut les entités physiques, les documents, les concepts abstraits, les nombres et les chaînes de caractères (w3c, 2014). Nous distinguons les ressources informationnelles qui sont directement accessibles sur le Web (documents, données, images, etc.) des ressources non informationnelles qui ne peuvent être appréhendées qu'au travers de représentations (phénomènes topographiques du monde réel, concepts abstraits, personnes, etc.). On appelle une « **ressource géoréférencée** » toute ressource qui est dotée dans sa description d'une référence spatiale. Une référence spatiale est une description d'une localisation dans le monde réel (ISO 19112, 2003). Elle peut être directe ou indirecte.

Les références spatiales directes décrivent d'une manière quantitative les caractéristiques spatiales des entités géographiques du monde réel telles que la localisation, la forme, l'orientation ou la taille. Elles peuvent être représentées au niveau des descriptions des ressources par des géométries (points, polylignes ou polygones) ou plus simplement des coordonnées, et moins souvent par une cellule dans une grille de référence. Ce travail de thèse porte principalement sur des ressources associées à des géométries ou des coordonnées.[1]

Les références spatiales indirectes, appelées « identifiant géographique (*geographic identifier*) » (ISO 19112, 2003), décrivent d'une manière implicite la localisation dans le monde réel. Il peut s'agir par exemple d'une adresse postale, d'un code postal, ou d'un nom de lieu. Cependant, pour être mieux exploitables par les systèmes d'information géographiques ou les applications exploitant des informations de localisation, les références spatiales indirectes peuvent être géocodées, *c.-à-d.* être traduites en références spatiales directes grâce à des services de géocodage ou des gazetiers. Les gazetiers sont des répertoires d'identifiants géographiques qui décrivent des localisations avec leurs références spatiales directes.

Nous nous intéressons dans ce chapitre à l'utilisation des références spatiales sur le Web de données. En effet, la géométrie joue un rôle central dans le domaine de données géographiques. Une base de données géographique vectorielle fournit une vision abstraite des phénomènes du monde réel qui réduit sa complexité. Les entités géographiques sont représentées dans des bases de données géographiques vectorielles par des objets, chacun décrit par un ensemble d'attributs qui le qualifient. La géométrie constitue l'attribut principal dans la description des objets car elle représente des caractéristiques telles que la localisation, la forme, l'orientation ou la longueur.

[1] Dans le cadre cette thèse les termes « référence spatial directe » et « géométrie » sont employés comme synonymes en ce qui concerne les ressource géoréférencées dans le Web

Dans le présent chapitre, nous introduisons tout d'abord les grands concepts du Web de données. Nous mettons ensuite l'accent sur la place de l'information géographique sur le Web de données. Nous présentons les différents efforts de standardisation de la représentation des données géographiques ainsi que les propositions de la représentation des références géographiques sur le Web de données. Nous présentons ensuite les enjeux et les verrous liés aux données géoréférencées sur le Web de données, notamment concernant leur interconnexion. Nous nous intéressons ensuite à l'exploitation des données géoréférencées publiées sur le Web, en particulier dans des applications de visualisation cartographique. Nous terminons par les objectifs de cette thèse.

1.1 Web sémantique et Web de données

Le Web sémantique fournit une plateforme commune qui permet aux données d'être publiées, partagées et réutilisées entre différentes applications, entreprises et communautés. C'est un effort collaboratif mené par le W3C avec la participation de nombreux partenaires de la recherche et de l'industrie (W3C, 2001). Le Web de données représente une concrétisation de la bonne utilisation des technologies qui fondent le Web sémantique. En effet, dans la lignée de son modèle du WWW (World Wide Web), Tim Berners-Lee (Berners-Lee, 2006) propose un modèle de données liées (Linked Data) qui repose sur les mêmes principes de base:

- Utiliser les URIs (*Uniform Resource Identifier*) pour nommer les choses, plus précisément des URIs HTTP qui permettent de consulter ces noms.
- Quand un URI est consulté, renvoyer des données (structurées) sur ce qu'il représente.
- Inclure dans ces données des liens vers d'autres URIs pour permettre de découvrir plus de données.

Ce modèle de donnés liée a permis d'étendre le Web des « documents » (Berners-Lee, 1989) en un Web de « données ». Tim Berners-Lee (Berners-Lee, 2006) propose également un ensemble de bonnes pratiques, qui repose sur un schéma de déploiement à 5 étoiles, agissant comme un système de notation qui vise à encourager les fournisseurs des données à adopter ce modèle des dans le but de créer un Web de données ouvertes et liées (Linked Open Data). Selon ce schéma, des données publiées sur le Web obtiennent une étoile si elles sont publiées sous une licence ouverte. Cela peut être dans n'importe quel format de fichier lisible (*ex.* des fichiers PDF). Une deuxième étoile leur est attribuée si elles sont dans un format structuré quelconque. Ceci permet aux données d'être traitées, transformées ou utilisées dans des calculs (*ex.* des fichiers EXCEL). La troisième étoile est obtenue si le format structuré utilisé pour représenter les données est non propriétaire. Ceci garantit une indépendance vis à vis des logiciels propriétaires pour traiter les données (*ex.* fichier CSV). Les données obtiennent une quatrième étoile si elles sont publiées dans les formats standards du Web sémantique, *c.-à-d.* utiliser les URIs pour identifier les choses et le modèle RDF (*Resource Description Framework*) pour structurer les données. Des données publiées sont considérées comme ayant cinq étoiles si, en plus de respecter les quatre premières conditions, elles sont liées à d'autres sources de données du Web. En plus de faciliter la découverte de données, les liens créés vers d'autres sources de données apportent une plus-value aux données publiées et rajoutent plus de possibilités d'applications du point de vue de l'utilisateur de ces données. Ce schéma d'implémentation concrétise l'utilisation des technologies du Web sémantique dans la création d'un espace global de sources de données interconnectées nommé le nuage des données ouvertes liées (Linked open data cloud).

La pile des standards du Web sémantique (voir Figure 1.1), mise au point par le W3C, fournit un modèle complet pour identifier, structurer, représenter, interroger, consulter, documenter et raisonner sur les données. L'utilisation des URIs fournit une identification universelle pour ces ressources. Une URI est « une séquence compacte de caractères qui identifie une ressource abstraite ou physique » (Archer, 2013). L'un des avantages des URIs est qu'ils peuvent être déréférençables, c'est-à-dire qu'ils permettent d'obtenir une représentation de la ressource identifiée. Le déréférencement permet, grâce à des mécanismes de négociation de contenu, d'obtenir l'accès à des descriptions de ces ressources compréhensibles en même temps par l'homme et par les machines.

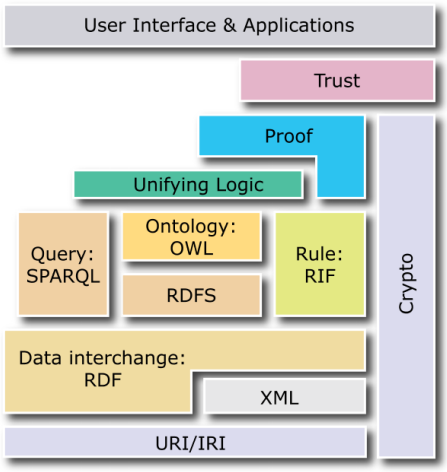


Figure 1.1 La pile des technologies du Web sémantique (Semantic Web layercake diagram « https://www.w3.org/2007/03/layerCake.png »)

Le modèle RDF (w3c, 2014) permet une représentation des données sous forme de graphe orienté et étiqueté. Dans ce modèle, les données sont structurées sous forme de triplets « sujet, prédicat, objet » (voir Figure 1.2). Un sujet dans un triplet RDF doit impérativement être une ressource identifiée par son URI. Le prédicat (ou la propriété) permet de décrire la relation entre le sujet et l'objet dans un triplet. Il est identifié par une URI et porte une sémantique décrite dans un vocabulaire. L'objet quant à lui peut être une valeur littérale typée, ou l'URI d'une autre ressource (qui peut être du même jeu de données comme d'un jeu de données distant). Plusieurs sérialisations existent pour encoder des données dans le modèle RDF: RDF/XML, Turtle, N-Triples, N-Quads, N3, JSON-LD, RDFa, etc.

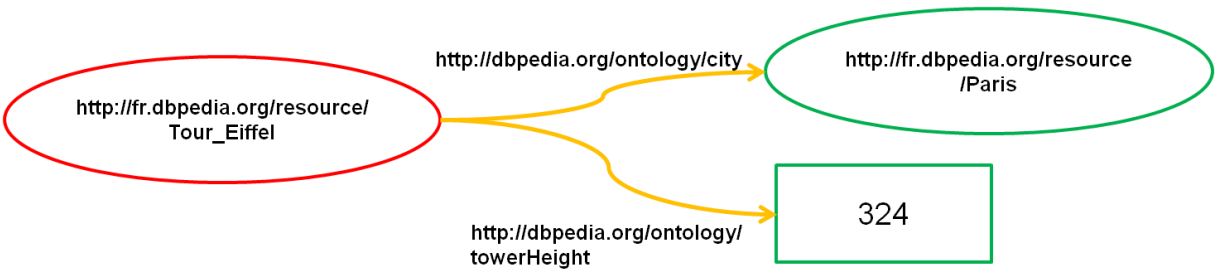


Figure 1.2 Exemple de triplets dans un graphe RDF : en rouge le sujet, en jaune les prédicats et en vert les objets.

Un graphe RDF est structuré et étiqueté conformément à des vocabulaires (ou ontologies) [2] définis au préalable et qui ont pour rôle de spécifier la sémantique des termes de ce graphe. Une ontologie est une « spécification explicite et formelle d'une conceptualisation partagée» (Gruber, 1995) qui, dans son sens informatique, fournit « un référentiel pour construire une base de connaissances permettant des inférences» et qui « inclut couramment une organisation hiérarchique des concepts pertinents et des relations qui existent entre ces concepts, ainsi que des règles et axiomes qui les contraignent » (Gandon et al., 2012). Pour représenter une ontologie, plusieurs langages, qui diffèrent en expressivité et en complexité, sont proposés, tels que SKOS[3], RDFS[4] ou OWL[5]. Les classes et propriétés d'un vocabulaire doivent, elles aussi, être identifiées par des conformément aux bonnes pratiques de définition et de publication d'ontologies proposées dans le but de garantir leur bonne réutilisation (Hyland et al. 2014).

Les données RDF peuvent être interrogées et manipulées grâce au langage de requêtes Sparql[6]. Il est à RDF ce que le langage SQL est aux bases de données relationnelles (Gandon et al. 2012). Ce langage est basé sur l'utilisation de patrons de triplets pour restreindre le résultat à obtenir à partir d'un Triple Store. Il définit dans sa version actuelle plusieurs opérateurs et fonctions sur les graphes et les valeurs.

L'émergence de ces différentes technologies pour la création d'un espace global de données liée incite de plus en plus de fournisseurs de données à publier leurs contenus selon le modèle des données liées. L'état actuel du nuage des données liées (voir Figure 1.3) montre la variété de natures de ces sources de données : elles peuvent être des sources génériques qui rassemblent des données de plusieurs domaines, comme elles peuvent être liées à une thématique bien précise (Andrejs et al., 2017). Les sources de données peuvent être d'origines différentes également : elles peuvent provenir de producteurs traditionnels ou d'institutions spécialisés, issues de plateformes participatives de collecte de données ou d'un mélange des deux.

1.2 La place centrale des ressources géoréférencées sur le Web de données

De nombreuses ressources du Web de données sont associées à une localisation dans l'espace, ou peuvent l'être du fait de leur nature. Ces ressources peuvent être des entités géographiques issues de base de données géographiques fournies par un producteur traditionnel de données, tel que une agence cartographique nationale (*ex.* données de L'Ordnance Survey[7], données de l'IGN Espagne[8], données l'IGN France[9]). Ces ressources peuvent également être dérivées de données issues de plateformes de saisie participative. C'est le cas par exemple de la source LinkedGeoData[10] issue de la

[2] Les termes vocabulaire et ontologie sont utilisée comme synonyme dans cette thèse (W3C, 2015).
[3] https://www.w3.org/2004/02/skos/
[4] https://www.w3.org/TR/rdf-schema/
[5] https://www.w3.org/OWL/
[6] https://www.w3.org/TR/sparql11-query/
[7] http://data.ordnancesurvey.co.uk/
[8] http://www.geo.linkeddata.es/
[9] http://data.ign.fr/
[10] http://linkedgeodata.org/

plateforme des données géographiques participative OpenStreetMap, ou les sources DBpedia[11] et Yago[12] issues en partie de Wikipedia. On peut trouver des ressources dotées de localisations dans des sources à origines hybrides. C'est le cas de GeoNames qui intègre des données saisies collaborativement en plus des données fournies par une multitude de producteurs de données géographiques.

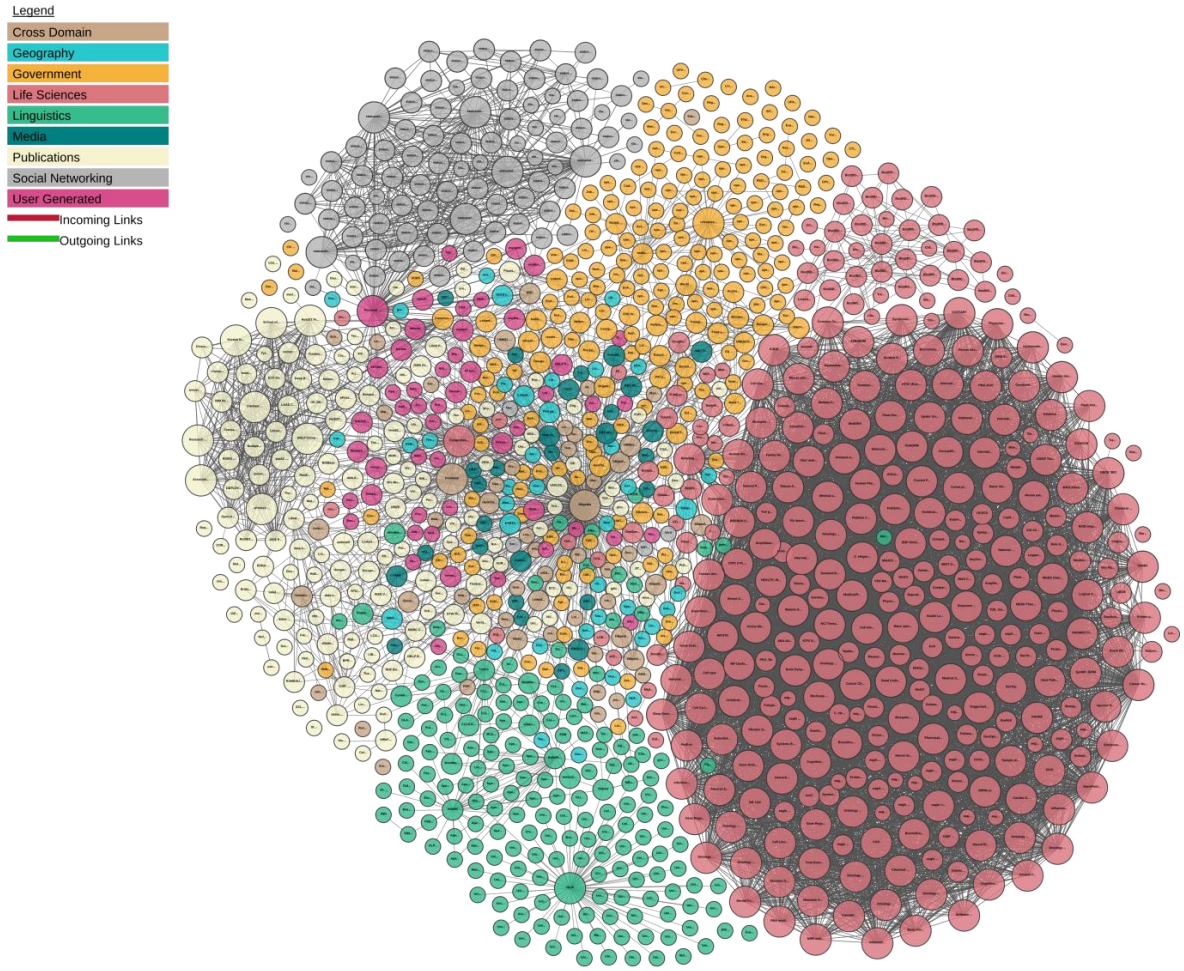


Figure 1.3 Diagramme du nuage de données ouvertes liées (*Linking Open Data cloud diagram)* 2017 (Andrejs et al., 2017)

Dans le nuage du Web de données, l'information spatiale constitue l'une des catégories de sources de données les plus importantes à la fois en termes de volume, mais également en termes de place au sein du nuage de données. Jusqu'en 2011 on pouvait identifier jusqu'à 31 sources qui contenaient plus que 6 milliards de triplets, ce qui représentait 19.43 % du contenu du nuage des Web de données selon le recensement « *State of the LOD Cloud*[13] ». Actuellement, le recensement du « *Mannheim Linked Data Catalog*[14] » compte 89 sources de nature géographique. L'importance de l'information géographique peut être égalemnt perçue par la taille de sources de données

[11] http://dbpedia.org/
[12] http://www.yago-knowledge.org/
[13] http://lod-cloud.net/state/state_2011/
[14] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/

géographiques. La source LinkedGeoData comprend 20 milliards de triplets pour décrire plus de 41 millions d' géographiques[15]. La source GeoNames, quant à elle, contient plus de 9 millions d'entités géographiques. L'utilisation de références spatiales pour géolocaliser des ressources est répandue également dans les autres catégories de sources de données. Par exemple, DBpedia, l'une des plus grandes sources de données généralistes sur le Web de données, contient dans sa version anglaise[16] des descriptions de plusde 1,1 million de ressources de type geo:SpatialThing, schema[17]:Place, dbo[18]:Place ou dbo:Location. Plus de 970000 de ces ressources sont géoréférencées par l'une ou plusieurs des propriétés spatiales suivantes : georss:point, wgs84:lon et wgs84:lat, ou wgs84:geometry. En outre, les interconnexions qui existent entre les sources de données géographiques et les sources de données des autres catégories qu'on peut constater sur le nuage des données ouvertes liées ont tendance à témoigner de l'importance et de la centralité de l'information géographique dans ce graphe de données.

Les géométries constituent l'information principale dans les bases de données géographiques vectorielles. Bien qu'elles ne jouissent pas forcément de la même importance dans la description des ressources du Web de données, les références spatiales de façon générale et les géométries particulièrement restent d'une grande utilité dans le traitement, le liage et l'exploration des données sur le Web.

L'un des bénéfices les plus importants que les références spatiales peuvent apporter réside dans la possibilité de les utiliser à des fins de mise en correspondance et de liage de données. Nous revenons sur ce point dans la partie 1.4.

Les références spatiales ajoutent aussi une dimension de recherche et de découverte de données grâce aux requêtes spatiales. En effet, elles permettent d'extraire des informations sur les ressources géoréférencées grâce à leurs caractéristiques spatiales (localisation, forme, taille, orientation, etc.) ainsi que découvrir les relations topologiques implicites qui peuvent exister entre elles.

Les informations portées per les géométries peuvent être croisées avec d'autres propriétés thématiques qui décrivent les ressources afin de dériver de nouvelles informations. Par exemple, un calcul de densité peut être réalisé à partir de la superficie d'une géométrie polygonale et d'une valeur de stock. Ceci permet d'enrichir et valoriser les données lors de leur publication.

En effet, la visualisation d'une ressource sur une carte, grâce à sa référence spatiale, est un moyen intuitif d'appréhender et de mieux comprendre le contexte de cette ressource (voir *ex.* Lodlive[19] qui permet entre autres de visualiser des données liées sur une carte dans la Figure 1.4). La visualisation cartographique d'une source de données, par le biais des références spatiales associées aux ressources qu'elle regroupe représente un moyen efficace pour explorer et même analyser visuellement les données (*ex.* le site[20] de visualisation cartographique des données de l'IGN et de

[15] Vérifié en janvier 2017
[16] Chiffres calculés en janvier 2017
[17] http://schema.org/
[18] http://dbpedia.org/ontology/
[19] http://en.lodlive.it/
[20] http://www.geo.linkeddata.es/browser.html

l'institut national de statistiques d'Espagne publiées en RDF). Nous revenons cet aspect dans la partie 1.5.

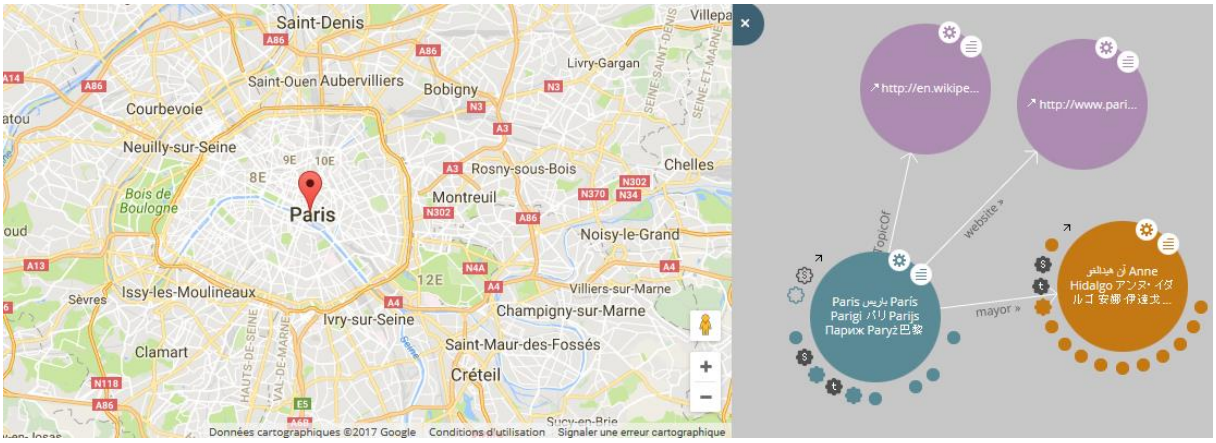


Figure 1.4 Extrait de la visualisation de la ressource http://dbpedia.org/resource/Paris dans Lodlive

1.3 Vers une représentation standard de données géoréférencées sur le Web de données

L'utilisation de toponymes est le moyen le plus intuitif pour associer une ressource à une localisation géographique. Cependant, un toponyme présenté sous la forme d'une chaîne de caractères ne permet pas une grande interopérabilité entre les sources de données : un même lieu peut être nommé différemment d'une source à une autre. Pour garantir plus d'interopérabilité, un moyen efficace revient l'URI d'une ressource qui décrit le lieu en question. Pour compléter la description de cette ressource avec une référence spatiale directe, la solution la plus simple est de lui rajouter des coordonnées géographiques. Si ce modèle de référence spatiale présente l'avantage d'être simple à comprendre est à intégrer à de données RDF, il ne permet pas, en revanche, de bénéficier des mêmes possibilités de représentation, manipulation et visualisation de données géoréférencées que celles offertes par les logiciels SIG (Système d'Information Géographique). L'intégration des normes et standards de la géomatique à ceux proposés dans le cadre du Web de données est donc primordiale pour bénéficier du potentiel de l'information géoréférencée. Nous présentons donc les différents efforts de standardisations proposés dans le domaine de l'information géographique pour la représentation la documentation et la manipulation de données géographiques vectorielles. Nous présentons ensuite les vocabulaires du Web de données proposés pour la représentation des références spatiales directes. Nous terminons par la description de la convergence des deux domaines à travers les efforts de standardisation communs du W3C et de l'OGC.

1.3.1 Des standards pour la représentation de l'information géographique

L'information géographique et sa représentation sont au centre des efforts de normalisation du comité technique TC 211 de l'organisation internationale de normalisation ISO (*International Standardisation Organisation*), et de standardisation du consortium OGC *(Open Geospatial Consortium*). Leurs travaux portent principalement sur la proposition de modèles pour la création, la représentation, la publication et l'utilisation de l'information géographique.

La norme ISO 19101 (Geographic information -- Reference model) fournit un modèle de référence de base qui offre une ligne directrice à un ensemble de normes pour la représentation de l'information géographique vectorielle. Elle introduit le concept de **Feature** comme étant une entité qui représente «une abstraction d'un phénomène du monde réel » et distingue donc les classes d'entités « **Feature Type** » des instances d'entités « **Feature Instance** ». Les caractéristiques de ces entités sont décrites par un ensemble d' « **attributs** ». Dans ce cadre, plusieurs autres normes ont été proposées. La norme ISO 19109 (Geographic information -- Rules for application schema) « définit des règles relatives à la création et la documentation de schémas d'application, y compris les principes de définition des entités ». Elle explique donc comment élaborer un schéma conceptuel d'entités géographiques et de leurs attributs afin de fournir une représentation d'un univers du discours. La norme ISO 19109 décrit les différents types d'attributs des entités géographiques, y compris l'attribut **GF_SpatialAttribute** qui représente les caractéristiques spatiales de ces entités géographiques, à savoir la géométrie **GM_Object** et la topologie **TP_Object**. La norme ISO 19107 (Geographic information -- Spatial schema) de son côté fournit un schéma conceptuel qui décrit en détail la classe des géométries **GM_Object** et la classe des topologies **TP_Object**. Selon cette norme, une géométrie **GM_Object** peut être une simple primitive géométrique, un objet complexe composé de plusieurs primitives ou un objet agrégeant plusieurs géométries. Afin de favoriser encore plus l'interopérabilité au niveau syntaxique, la norme ISO 19136 (Geographic information -- Geography Markup Language (GML)) présente un ensemble de schémas XML qui permettent de représenter les données géographiques et leurs schémas. Pour cela, cette norme propose l'utilisation du standard d'encodage OCG GML 3.2.1.

Dans la série ISO 19100, plusieurs autres normes ont été proposées autour de la représentation, la documentation et la publication des données géographiques. La norme ISO 19110 (Geographic information -- Methodology for feature cataloguing) fournit par exemple une méthodologie qui vise à améliorer l'exploitabilité des données d'une base de données géographique par la création de catalogues qui décrivent les classes d'entités, leurs opérations, leurs attributs et les relations qui peuvent exister entre les entités. La norme ISO 19131 (Geographic information -- Data product specifications) définit un modèle conceptuel qui permet de décrire les différentes spécifications d'une base de données géographique. Un autre exemple est la norme ISO 19115 (Geographic information -- Metadata) qui définit un modèle pour la représentation des différentes métadonnées d'une base de données géographique. Dans le même cadre, la norme ISO 19157 (Geographic information -- Data quality) définit un modèle conceptuel pour la représentation des métadonnées spécifiquement liée à la qualité des données géographiques. La norme ISO 19139 (Geographic information -- Metadata implementation specification) décrit comment encoder en XML les métadonnées conformes au modèle ISO 19115.

L'implémentation des résultats de normalisation et de standardisation décrits ici dans des outils SIG a permis de réduire les problèmes liée à l'hétérogénéité de la syntaxe et des modèles conceptuels des données géographiques. Ces normes et standards ont été par exemple utilisés dans le cadre de la directive INSPIRE[21]. Cette directive adoptée par le Parlement et le Conseil de l'Union Européenne vise à créer des infrastructures d'information géographique pour ses états membres. Ces infrastructures servent à regrouper, publier et cataloguer d'une manière cohérente des données issues de plusieurs

[21] http://inspire.ec.europa.eu/

sources pour faciliter leur utilisation. Dans ce cadre, des normes telles que l'ISO 19109, 19115 et 19131 sont utilisées dans la définition des schémas de données, de leurs métadonnées, ainsi que de leurs spécifications.

1.3.2 Des vocabulaires pour la représentation des références spatiales directes sur le Web de données

Depuis les débuts du Web de données, de nombreux vocabulaires ont été proposés pour répondre au besoin de représenter les entités géographiques et leurs références spatiales. Dans ce contexte, (Atemezing et Troncy, 2012) identifie deux types de vocabulaires : ceux qui permettent de représenter les types et les propriétés des entités géographiques et ceux qui permettent plus spécifiquement de représenter les géométries. (Salas et Harth, 2011) et (Atemezing et Troncy, 2012) recensent différentes manières possibles pour représenter la géométrie. Nous distinguons deux types de représentations possibles : littérale et structurée.

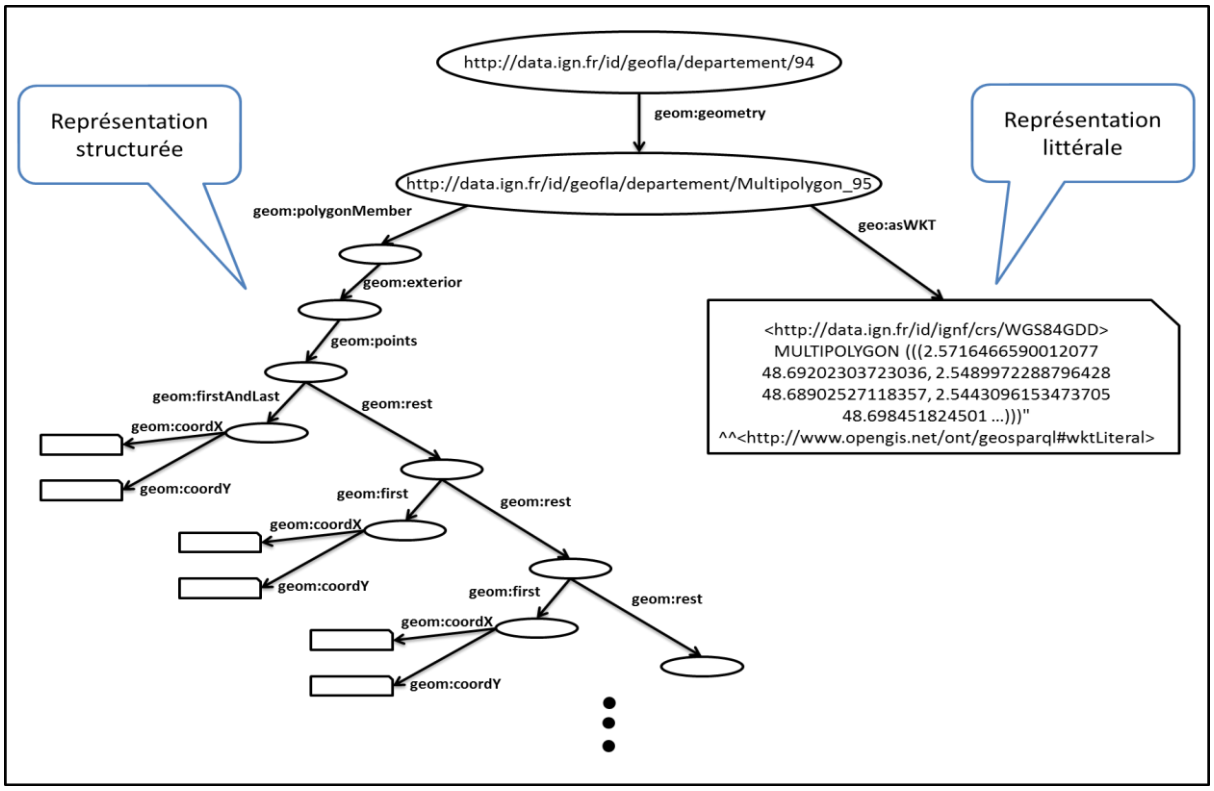


Figure 1.5. Exemple de représentations littérale et structurée de la géométrie de la ressource qui décrit le département du Val-de-Marne http://data.ign.fr/id/geofla/departement/94

La représentation littérale (voir *ex.* Figure 1.5 à droite) consiste à donner une description de la géométrie sous forme de chaîne de caractères. Un exemple de cela est la représentation des étendues spatiales (indiquées par la propriété osgo[22]:extent) des données de l'Ordnance Survey[23] dont la géométrie est introduite par la propriété osgo:asGML qui prend comme valeur un littéral XML qui contient la description de la géométrie selon le format d'encodage GML. Ce principe de

[22] http://data.ordnancesurvey.co.uk/ontology/geometry/
[23] http://data.ordnancesurvey.co.uk

sérialisation de géométries sous forme littérale est le même proposé par l'ontologie du standard GeoSparql[24] de l'OGC. Selon cette ontologie, les entités de type « Feature » sont liées à des géométries de type « Geometry » par une propriété geo[25]:hasGeometry. Les géométries peuvent être décrites par deux propriétés: geo:asGML et geo:asWKT, qui prennent comme valeurs des chaînes de caractères GML et WKT (Well Known Text[26]). Dans ces modes de sérialisation, la chaîne de caractères représentant la géométrie inclut son type de primitive géométrique et son système de coordonnées. Un autre exemple de représentation littérale est l'encodage simple des géométries par GeoRSS[27], conçu principalement pour les flux RSS. Dans ce cas, les géométries sont représentées par une liste de coordonnées représentant le(s) point(s) qui les constituent. Le type de primitive de la géométrie est, lui, indiqué par le prédicat utilisé : georss[28]:point, georss:box, georss:line ou georss:polygon. Son système de coordonnées de référence, en revanche, n'est pas précisé.

La représentation structurée (*ex.* Figure 1.5 à gauche) des géométries consiste à les décrire par un ensemble de sous éléments structurants allant jusqu'au niveau des coordonnées. Dans ce mode de représentation, les coordonnées de points sont généralement fournies sous forme de deux propriétés de longitude et de latitude ou d'easting et de northing selon le système de coordonnées utilisé. Les propriétés les plus utilisées pour cela sont wgs84[29]:long et wgs84:lat proposées par le W3C pour représenter des valeurs de longitude et de latitude dans le système de coordonnées de référence WGS84. On peut également mentionner les propriétés os[30]:easting et os:northing utilisées par l'Ordnance Survey pour la représentation des coordonnées planes. Cette forme de localisation est présente aussi dans les sources de données qui ne sont pas spécifiquement de nature géographique telles que DBpedia qui utilise les propriétés prop-fr[31]:longitude et prop-fr:latitude ou Yago qui utilise les propriétés yago[32]:hasLongitude et yago:hasLatitude. Une autre forme de représentation structurée est la localisation par le rectangle englobant de la ressource géoréférencée. Ce rectangle peut être représenté par exemple sous forme de quatre propriétés (longitudes minimum et maximum, latitudes minimum et maximum) comme dans le cas de l'ontologie géopolitique de la FAO[33]. On note cependant que dans ces trois derniers exemples les propriétés de longitude et de latitude sont liées directement à la ressource géoréférencée sans l'utilisation d'une ressource « géométrie » intermédiaire. Ce choix de représentation de coordonnées a de fortes limites, car il contraint la ressource à être liée à une seule représentation géométrique de sa localisation. Dans le cas où la géométrie est plus complexe qu'un simple point, certaines sources utilisent une structure sous forme de séquence de points pour la représenter. Les ontologies LinkedGeoData (Auer et al., 2009) ou Geo.linkeddata (De León et al., 2010) utilisent cette forme de présentation. Dans le cas de LinkedGeoData, en plus d'être décrites par sérialisation littérale par la propriété geo:asWKT, les géométries linéaires sont décrites par une ressource de séquence liée à toutes les ressources de type « nœud » qui représentent les points constituant la ligne. La ressource

[24] http://www.opengeospatial.org/standards/geosparql
[25] http://www.opengis.net/ont/geosparql#
[26] http://www.opengeospatial.org/standards/wkt-crs
[27] http://www.georss.org/
[28] http://www.georss.org/georss/
[29] https://www.w3.org/2003/01/geo/wgs84_pos
[30] http://data.ordnancesurvey.co.uk/ontology/spatialrelations/
[31] http://fr.dbpedia.org/property/
[32] http://yago-knowledge.org/resource/
[33] http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource/

de séquence est de type rdf:Seq[34]. Elle est donc liée à ces nœuds par des prédicats de type rdfs:ContainerMembershipProperty[35] dont les instances prennent la forme rdf:_n telle que n représente la position du nœud dans la séquence. Une structuration géométrique semblable est proposée par la plateforme Geo.linkeddata pour la représentation de ses géométries. Chaque géométrie de type ligne ou polygone est liée aux points qui la constituent avec une propriété dcterms[36]:hasPart. Chaque point spécifie son ordre dans la géométrie qui le contient par un prédicat geoes[37]:order.

La typologie des géométries varie aussi d'un vocabulaire à un autre. Par exemple, l'ontologie GeoSparql propose une seule classe générale « Geometry » pour la représentation des géométries, alors que l'ontologie « Simple Features Geometry (sf) [38] » propose des spécialisations de cette classe pour préciser s'il s'agit d'un point, d'une polyligne, d'un polygone, d'une multi-polyligne, d'un multi-polygone, etc. L'ontologie « NeoGeo (ngeo)[39]» spécialise de la même manière la classe «ngeo:Geometry», mais en décrivant aussi les relations de structuration qui peuvent exister entre ces classes (*ex.* la propriété ngeo:polygonMember qui permet de lier un multi-polygone aux polygones qui le constituent).

L'ontologie geom[40] de l'IGN (Troncy et al., 2014; Hamdi et al., 2014) regroupe plusieurs des avantages des vocabulaires cités ci-dessus. Dans cette ontologie, la classe geom[41]:Geometry spécialise les classes sf:Geometry et ngeo:Geometry. Ce vocabulaire permet de représenter des géométries, à la fois sous forme structurée conformément au modèle standard « Simple Feature Access » et sous forme littérale conformément au standard GeoSparql. Pour cela, elle implémente des structures de liste (rdf:list[42]) pour représenter les listes ordonnées des points composant le contour de chaque géométrie. Cette ontologie permet de spécifier le système de coordonnées au niveau de chaque géométrie grâce à une propriété geom:crs, qui prend comme valeur l'URI identifiant son système de coordonnées. Cet URI peut provenir du registre[43] IGNF qui contient les descriptions, structurées selon l'ontologie ignf[44], des systèmes de coordonnées de référence définis et maintenus par l'IGN.

1.3.3 Un effort de standardisation commun W3C - OGC

L'existence de vocabulaires pour la description des entités spatiales et de leurs géométries peut être perçue comme un moyen de réduire l'hétérogénéité entre les données. Effectivement, c'est le cas aux niveaux syntaxique et schématique, voire sémantique, dans la mesure où cela permet de publier les données, selon le même vocabulaire ou selon des vocabulaires alignés, dans le même modèle

[34] https://www.w3.org/TR/rdf-schema/#ch_seq
[35] https://www.w3.org/TR/rdf-schema/#ch_containermembershipproperty
[36] http://purl.org/dc/terms/
[37] http://geo.linkeddata.es/
[38] http://www.opengis.net/ont/sf
[39] http://geovocab.org/geometry#
[40] http://data.ign.fr/def/geometrie
[41] http://data.ign.fr/def/geometrie#
[42] https://www.w3.org/TR/rdf-schema/#ch_list
[43] http://data.ign.fr/id/ignf/
[44] http://data.ign.fr/def/ignf

(RDF). Cependant, il n'existe pas de consensus sur l'(les) ontologie(s) à utiliser. Les bonnes pratiques de définition et de publication des ontologies (Hyland et al., 2014) ne sont pas forcément respectées par toutes les sources de données. Par exemple, certains jeux de données utilisent leurs propres vocabulaires qui redéfinissent des propriétés ou des classes qui existent par ailleurs sans y faire référence : selon les statistiques de LODstats[45], il existe au moins 40 paires de propriétés (autres que wgs84:long et wgs84:lat) utilisées par au moins 45 jeux de données pour représenter les coordonnées en longitude et latitude. Les données géoréférencées du Web ne sont pas toujours facilement utilisables. Les données sont souvent publiées sans les métadonnées qui permettent de les identifier, les découvrir, connaître leur qualité ou les comprendre. Les données géoréférencées publiées sur le Web soulèvent donc des problèmes de découverte, d'accessibilité et d'interopérabilité. Ces mêmes problèmes sont ceux qui ont été particulièrement visés par les propositions des normes ISO 19100 et les standards OGC, qui s'avèrent complexes à implémenter dans leur intégralité. Augmenter la fiabilité et l'utilisabilité des données nécessite donc de trouver le meilleur compromis entre les bénéfices apportés par ces normes et standards d'un côté, et d'un autre l'accessibilité et la simplicité d'utilisation prônées par les standards du W3C.

L'enjeu majeur que représente l'information géographique en général sur le Web a suscité un rapprochement entre le W3C et l'OGC. Les deux organismes de standardisation ont créé en 2014 un groupe de travail conjoint nommé « *Spatial Data on the Web Working Group*[46]». Ce groupe a comme mission principale de déterminer les meilleurs moyens pour intégrer l'information géographique sur le Web de données, en définissant un ensemble de bonnes pratiques (SDWBP)[47] qui viennent pour compléter la recommandation des bonnes pratiques de publication de données sur le Web (DWBP)[48]. Ceci permet de compléter la standardisation des propositions existantes les plus répandues sur le Web pour la représentation de l'information géographique. Dans leur état actuel[49], les bonnes pratiques proposées par ce groupe de travail couvrent principalement les aspects liés à l'information spatiale suivants :

- Les métadonnées spatiales
- La qualité spatiale des données (notamment la précision de position)
- La gestion des versions des données spatiales
- Les identifiants des données spatiales
- L'utilisation des vocabulaires spatiaux (la description des références spatiales, la sémantique des données spatiales et l'aspect temporel des données spatiales)
- L'accès aux données spatiales
- Le liage des données spatiales
- La gestion des jeux de données spatiales volumineux

Le document des bonnes pratiques précise qu'il prend comme point de départ les méthodes utilisées habituellement dans les SDIs (*Spatial Data Infrastructures* ou infrastructures des données géographiques), pour étendre leur application aux données liées. Les liens entre ressources étant

[45] http://stats.lod2.eu/
[46] https://www.w3.org/2015/spatial/wiki/Main_Page
[47] Spatial Data on the Web Best Practices : https://www.w3.org/TR/sdw-bp/
[48] Data on the Web Best Practices : https://www.w3.org/TR/dwbp/
[49] W3C Working Group Note 30 March 2017: https://www.w3.org/TR/2017/NOTE-sdw-bp-20170330/

l'élément principal du modèle des données liées, deux des bonnes pratiques qui sont définies dans ce document sont consacrées à la publication et l'utilisation des liens associés aux données spatiales. La publication des liens entre entités spatiales à l'intérieur d'une même source ou entre différentes sources nécessite en premier lieu leur découverte. Le document ne discute pas précisément des moyens disponibles pour découvrir des liens entre entités spatiales. Néanmoins, les bonnes pratiques d'accessibilité, de choix de vocabulaire, d'inclusion de métadonnées, etc., valorisent amplement le rôle que les géométries peuvent jouer dans les différentes applications recensées dans la section 1.2, y compris l'identification des liens entre les différentes ressources, tâche à laquelle nous nous intéressons principalement dans ce travail.
Nous présentons dans la suite le rôle primordial que les références spatiales peuvent jouer dans l'interconnexion de ressources géoréférencées, et les défis à relever lors de leur utilisation pour la découverte des liens.

1.4 Lier les ressources géoréférencées du Web : enjeux et verrous

L'interconnexion (ou le liage) des données est le processus qui fournit les liens qui constituent les fils de la toile qu'est le Web de données. L'interconnexion consiste à identifier des relations de correspondance entre les ressources et le cas échéant, à les matérialiser sous la forme de liens. (Ferrara et al. 2013) la définit plus précisément comme étant « la tâche de déterminer si les descriptions de deux ressources peuvent être liées du fait qu'elles fassent référence à la même entité du monde réel dans un domaine spécifique, ou qu'une sorte de relation existe entre elles ». Les descriptions des ressources peuvent contenir des identifiants universellement acceptés et réutilisés (*ex.* les codes ISBN et ISSN pour les publications, les code EAN et EPC pour les produits, etc.). Ceci permet d'établir facilement un lien d'équivalence entre les ressources. En l'absence de ce genre d'identifiants, une similarité doit être calculée entre ces ressources pour pouvoir déterminer s'il est pertinent ou pas de créer un lien entre elles (Bizer et al., 2009). Ceci rejoint des travaux existants de la communauté des bases de données sur le couplage d'enregistrement en fichiers et en base de données (record linkage) (Winkler, 2006 ; Christen, 2012), la résolution d'entités et la détection de doublons (entity resolution, duplicate detection) (Christen, 2012). Ceci rejoint également les travaux sur l'alignement d'ontologies (ontology matching) (Euzenat et Shvaiko, 2007), sauf qu'il s'agit dans le cas de l'interconnexion d'établir des liens entre les instances de données (les ressources) et non entre les éléments des schémas de données (les ontologies).

1.4.1 Lier les ressources grâce à leurs références spatiales : bénéfices et applications

Dans le domaine des données géographiques, L'appariement est le processus qui vise à mettre en correspondance des objets géographiques de bases de données géographiques hétérogènes qui représentent un même phénomène du monde réel (Devogele, 1997 ; Brando, 2013). L'appariement représente le cœur de plusieurs applications liées aux données géographiques, telles que l'évaluation de qualité, la gestion des mises à jour, le recalage des données ou encore l'intégration de données (Olteanu,2008). Dans un processus d'appariement de données géographiques, on cherche à mettre en correspondance des objets en les comparant selon un ou plusieurs critères. La géométrie est généralement le critère principal utilisé dans l'appariement de données géographiques du fait de son rôle prépondérant dans les bases de données géographiques. Les géométries sont généralement utilisées dans l'appariement en suivant l'intuition selon laquelle si deux géométries de deux bases de

données géographiques différentes sont spatialement proches, alors elles sont très susceptibles de représenter la même entité géographique du monde réel. L'état de l'art concernant l'appariement de données géographiques comprend de nombreuses approches, techniques, mesures de distance et outils qui sont dédiés à l'utilisation des références spatiales comme critère de mise en correspondance des entités géographiques. Cet état de l'art conséquent (sur lequel nous reviendrons dans la section 2.1) montre l'apport qu'une utilisation adéquate des références spatiales peut offrir à la tâche de mise en correspondance des ressources du Web.

De nombreuses approches d'interconnexion automatiques ou semi-automatiques (*cf.* section 2.2) s'appuient sur la ressemblance des « clés » (Pernelle et al., 2013). Une clé est une propriété qui permet d'identifier d'une manière unique une ressource (Atencia et al., 2012). Plusieurs approches ont été proposées pour la détection de ces à des fins d'interconnexion de données (Pernelle et al., 2013) (Atencia et al., 2012). (Symeonidou et al., 2014) introduit la notion de « quasi-clé » qui représente les propriétés qu'on ne peut pas considérer comme clés à cause d'une minorité de ressources qui ne garantissent pas la notion d'unicité. La combinaison de plusieurs quasi-clés permet cependant d'identifier les ressources d'une manière unique. La caractéristique principale d'une clé (ou d'une quasi-clé) demeure son caractère discriminant. Une référence spatiale directe peut donc très fortement constituer une des propriétés clés (ou quasi-clés) dans un jeu de données. En effet, dans un jeu de données qui contient des ressources géoréférencées, il est très peu probable que deux ressources distinctes partagent la même référence spatiale. Donc, les références spatiales sont un critère pertinent supplémentaire (si ce n'est l'unique dans certains cas) à prendre en compte dans un processus d'interconnexion.

Cependant, on peut supposer qu'une ressource du Web représentant une entité géographique ne soit pas forcément décrite par une référence spatiale directe. Ceci est dû au fait que les références spatiales directes dans le contexte du Web de données ne jouissent pas de la même importance que les géométries dans les bases des données géographiques. En revanche, la localisation de cette entité géographique pourra être décrite par une référence spatiale indirecte comme une adresse par exemple. On dispose d'un état de l'art considérable de techniques qui permettent d'associer des références spatiales directes aux ressources représentant des entités géographiques et localisées à l'aide d'identifiants spatiaux. Il s'agit principalement de techniques de géocodage ou de résolution d'entités spatiales nommées.

Le géocodage est le processus qui permet de transformer un référencement spatial descriptif (*c.-à-d.* indirect) en un référencement spatial quantitatif (*c.-à-d.* direct)(Goldberg et al., 2007). Il s'agit d'utiliser des données de référence qui relient références spatiales directes et indirectes comme les bases de données d'adresses ou de subdivisions administratives. De plus en plus de services Web de géocodage ou de jeux de données brutes d'adresses sont disponibles sur le Web. Citons par exemple le service de géocodage de Google[50], celui de l'IGN[51] ou celui[52] de la Base Adresse Nationale (BAN)[53]. Cette tâche qui était historiquement un processus très coûteux et lent (Goldberg et al., 2007), devient désormais de plus en plus facile grâce aux outils et données disponibles sur le Web.

[50] https://developers.google.com/maps/documentation/geocoding/intro
[51] https://mesadresses.ign.fr/
[52] http://adresse.data.gouv.fr/api/
[53] http://adresse.data.gouv.fr/

Dans une optique similaire, la résolution d'entités spatiales nommées est un domaine dont l'objectif général et d'associer une mention d'entité spatiale extraite d'un texte au toponyme correspondant dans une base de données géographique ou un gazetier. Le problème principal se pose quand, pour une mention d'entité spatiale, on trouve plusieurs toponymes candidats (Sehgal et al ., 2006). Le but dans ce cas est de désambiguïser cette entité spatiale nommée. L'état de l'art des approches et outils de ce domaine (voir (Moncla, 2015)) fournit une palette de moyens qui permettent d'associer à une mention d'entité spatiale un toponyme bien identifié, et lorsque cette information est disponible dans la base de référence, de lui associer une référence spatiale directe. Ces mêmes approches et outils peuvent être transposés dans le contexte du Web de données pour fournir, d'une manière non ambigüe, des références spatiales directes aux ressources représentant des entités géographiques. En outre, les différentes approches de résolution d'entités spatiales nommées peuvent inspirer les méthodes d'interconnexion de ressources du fait qu'elles partagent l'objectif de base de trouver des correspondances entre des ressources de sources de données hétérogènes. Les approches de résolution d'entités spatiales nommées se répartissent en trois grandes familles selon (Buscaldi, 2011). Une de ces familles d'approches s'appuie principalement sur l'utilisation des références spatiales pour la désambiguïsation des entités spatiales nommées se trouvant dans un même contexte (*ex.* texte, paragraphe, phrase, etc.). Ce genre d'approches se fonde sur l'hypothèse que des entités spatiales nommées mentionnées dans un même contexte textuel sont très susceptibles d'être spatialement proches. Ainsi, si les villes Paris, Vincennes et Saint-Mandé sont citées dans la même phrase, il y a de fortes chances que la mention de « Paris » désigne la capitale française et non son homonyme texane. Le géocodage et la résolution d'entités spatiales nommées permettent d'intégrer des ressources à localisation indirecte dans un même référentiel spatial que les ressources à références spatiales directes. Ceci permet d'élargir l'éventail des applications possibles sur les données en offrant la possibilité d'effectuer des analyses spatiales ou de développer des solutions de visualisation cartographique, etc. On peut donc s'inspirer de ces approches dans l'utilisation des références spatiales pour l'interconnexion des ressources du Web.

L'interconnexion vise à rechercher des correspondances entre les ressources du Web et de créer des liens entre les ressources identifiées comme homologues. Cette correspondance n'est pas forcément une relation d'équivalence, même si c'est le plus souvent le cas. En effet, (Ferrara et al., 2013) précise qu'il peut s'agir de rechercher des relations d'équivalence entre les ressources (généralement matérialisées par des liens owl:sameAs), comme il peut s'agir de relations qui portent autres sémantiques. Avoir des relations d'équivalence entre ressources géoréférencées permet d'obtenir des représentations géométriques multiples pour une même entité du monde réel. Ceci peut être très bénéfique pour la visualisation cartographique des ressources du Web en offrant des représentations géométriques adaptées aux différentes échelles de visualisation. Au-delà des relations d'équivalence, grâce aux références spatiales, le processus d'interconnexion peut être utilisé pour joindre les ressources spatialement. On peut ainsi joindre des ressources dont la description est riche en information thématique à des ressources dont la description géométrique est beaucoup plus détaillée. Par exemple, on peut lier une ressource décrivant une administration à la ressource qui décrit le bâtiment qui l'abrite, ou lier la ressource qui décrit un événement à la ressource qui décrit le lieu où il se déroule comme un bâtiment, un terrain de sport ou une ville. Croiser des ressources décrivant des entités géographiques à des ressources portant des informations thématiques ouvre la voie à des applications comme la création de cartes thématiques ou l'analyse thématique d'un espace. Grâce aux références spatiales, on peut également créer des

liens qui explicitent des relations spatiales, comme dans le travail de (Ahmed Sherif et al., 2017) qui permet d'expliciter les relations topologiques par interconnexion.

Les approches d'appariement des données de bases de données géographiques, comme celles d'interconnexion des données géoréférencées du le Web, se différencient les unes des autres par les défis auxquels elles font face : la taille des données, les limitations des ressources espace et temps ou l'hétérogénéité des données. Cette dernière constitue généralement le défi principal relevé par les différentes approches d'appariement de données géographiques. Nous exposons dans la suite les différents types d'hétérogénéités qui peuvent se présenter entre les jeux de données géographiques et nous montrons comment ces formes d'hétérogénéités se retrouvent dans le cas des sources du Web de données, non plus seulement entre jeux de données, mais également entre ressources au sein d'un même jeu de données. Nous reviendrons dans le chapitre 2 sur les différentes approches d'appariement et d'interconnexion de données géoréférencées du Web de données.

1.4.2 Hétérogénéités des références spatiales : des bases de données géographiques au Web de données

Pour comprendre les hétérogénéités qui existent entre les représentations des références spatiales des ressources géoréférencées du Web, nous devons comprendre celles qui existent déjà dans le domaine des bases de données géographiques. Une base de données géographique vectorielle fournit une représentation de l'espace géographique en un instant donné, partielle et non unique. Elle est créée selon une abstraction du monde réel qui réduit sa complexité (Girres, 2012) et qui dépend du point de vue adopté au préalable par le producteur de cette base (Fonseca et al., 2003). Des différents objectifs et points de vue adoptés lors de la production de deux bases de données géographiques résultent forcément des différences entre celles-ci.

Pour les bases de données géographiques, comme pour n'importe quel type de bases de données, on peut trouver trois types principaux d'hétérogénéité (Kavouras et Kokla, 2008 ; Euzenat et Shvaiko, 2007) : une hétérogénéité syntaxique (les modèles conceptuels des données sont différents), une hétérogénéité schématique (les schémas conceptuels des données utilisent des attributs différents pour une même classe), ou une hétérogénéité sémantique (*ex.* deux classes portent le même nom mais représente des phénomènes différents). Les différents efforts de normalisation et standardisation (ISO et OGC *cf*. 1.3) ont considérablement contribué à réduire l'hétérogénéité syntaxique et ont au moins permis, même partiellement, de mieux appréhender les différentes hétérogénéités qui peuvent exister entre les données géographiques. La sémantique, selon (Kavouras et Kokla, 2008), correspond à la relation entre les données et les phénomènes du monde réel qu'elles représentent. L'hétérogénéité sémantique peut être causée par une différence de classification des concepts due à des points de vue différents : une pharmacie peut être considérée comme un commerce selon un point de vue, ou comme un élément du système de santé selon un autre. Elle peut être aussi causée par un partitionnement différent de la réalité : *ex.* le concept « canal » peut avoir deux sous-concepts « irrigation » et « transport » dans une base, et deux autres sous-concepts « irrigation » et « drainage » dans une autre base. L'hétérogénéité sémantique peut être causée également par la différence des niveaux de détail des bases de données.

Une base de données géographique est décrite par une composante spatiale (géométrique) et une composante non spatiale (dite sémantique)(Ruas, 2002). D'un point de vue géométrique, le niveau

de détail des données géographiques est défini par la conjonction de « la résolution », « la précision géométrique » et « la granularité » des géométries présentes dans une base (Ruas, 2002; Touya et Brando, 2013). La résolution géométrique donne l'ordre de grandeur géométrique des phénomènes présents dans la base. Elle correspond en principe à la plus petite unité mesurable ou représentable. Cette notion est souvent applicable à des données maillées (Raster), car elle correspond originellement à la dimension représentée par un pixel. Cette notion demeure ambigüe pour les données vectorielles et doit préférablement être fournie avec la précision et la granularité géométrique (Ruas, 2002). La résolution peut être définie simplement par une échelle de raisonnement équivalente, dans la mesure où la façon dont on modélise le monde réel change selon cette échelle. La précision géométrique représente l'écart entre la position représentée dans la base et la position effective de l'entité géographique dans le monde réel. Comme cet écart n'est pas forcément uniforme pour toutes les géométries d'une base, elle est souvent représentée par la moyenne de ces écarts (Ruas, 2002). La granularité représente la taille de la plus petite forme représentée. Elle ne concerne pas la représentation en points. Elle est définie pour les lignes comme étant la taille du plus petit segment dans un virage et pour les polygones comme étant la taille du plus petit décrochement (Ruas, 2002).

Une base de données géographique est créée conformément à un ensemble de règles décrites dans des documents, qu'on appelle « spécifications », et qui reflètent le point de vue du producteur (Girres, 2012). Les spécifications sont définies de manière générale par la norme ISO8402 comme étant « le document qui prescrit les exigences auxquelles le produit ou le service doit se conformer » (ISO8402, 1994). On distingue deux types de spécifications concernant les bases de données géographiques (Girres, 2012) : « les spécifications de contenu » qui décrivent ce que la base de données géographique doit contenir, et « les spécifications de saisie » qui décrivent comment produire le contenu de la base de données en termes de techniques et de méthodes. Les spécifications d'une base de données géographique assurent l'homogénéité des données au sein de cette dernière. Elles regroupent l'ensemble de règles définissant le contenu de chaque classe de la base de données géographique (Abadie, 2012). Dans les spécifications des bases de données géographiques de l'IGN par exemple, chaque classe est décrite par une fiche qui la définit et précise son type de primitive géométrique. Elle décrit également l'ensemble des règles de sélection, selon des caractéristiques spatiales, des entités du monde réel concernées. Elle précise ensuite les règles qui définissent comment les entités concernées doivent être saisies : on parle alors de la« modélisation géométrique » des données. Cette dernière décrit comment la géométrie doit être constituée à partir l'entité réelle représentée. Par exemple, un bâtiment doit être représenté géométriquement par le contour du toît et non pas par celui des murs. La fiche se termine avec la liste des attributs qui servent à décrire les objets de la classe en question, leurs définitions, et leurs valeurs énumérées.

On constate donc que si deux bases de données sont définies pour deux niveaux de détail différents, ou selon des règles de sélection et de saisie différentes, ceci constitue une cause d'hétérogénéité entre leurs géométries. Selon (Devogele, 1997), ces différences constituent des « conflits d'intégration » qu'on doit identifier pour les prendre en compte quand on veut intégrer des bases de données géographiques. Il propose donc une taxonomie de ces conflits d'intégration. Selon cette taxonomie, les conflits peuvent être liés aux sources et origines des données, aux modèles et métadonnées des données, aux définitions des classes, aux structures utilisées pour représenter ces classes, aux descriptions sémantique et géométrique des entités géographiques, ou aux valeurs de

leurs attributs. Ces conflits sont liés les uns aux autres en sachant qu'un conflit peut en engendrer d'autres. Ainsi, (Sheeren, 2005) tire de cette taxonomie six principaux types de conflits différents. Le premier est le conflit de critères de sélection entre les bases de données. Il peut s'agir d'une différence de taille minimale qu'une entité doit avoir pour être sélectionnée au sein de deux bases de données géographiques différentes. Ceci peut causer une différence d'exhaustivité d'instances de classes (supposément homologues) entre les deux bases. Le deuxième conflit est celui de critères de décomposition. Un critère de décomposition précise à partir de quel seuil une entité ne doit pas être représentée par un seul objet, mais par plusieurs. Si ce seuil change pour deux classes homologues de deux bases différentes alors on peut s'attendre une différence de niveau de détail des instances de ces classes. Le troisième type de conflit est très important, car il s'agit des critères de la modélisation géométrique. La complexité de cette modélisation géométrique est généralement liée l'échelle équivalente de la base de données: on peut imaginer qu'un cours d'eau peut être représenté par un polygone qui représente sa surface dans une base destinée à être utilisée à grande échelle, alors qu'il peut être représenté par une ligne qui représente son axe dans une autre base destinée à être utilisée à petite échelle. Un quatrième conflit de résolution se manifeste quand les tailles des plus petits objets représentables sont différentes d'une base à une autre. Cela peut concerner la superficie minimum qu'une cour d'un bâtiment doit avoir pour être saisie. (Abadie, 2012) préfère nommer ce type de conflit « conflit de granularité ». Ceci nous amène au cinquième conflit, celui de granularité (Devogele, 1997), que (Abadie, 2012) préfère nommer « conflit de seuil de segmentation ». Il s'agit ici d'une différence dans des seuils de segmentation des données entre des classes homologues ayant les mêmes critères de décomposition. Par exemple, les routes peuvent être décomposées en tronçon selon leurs natures ou leurs nombres de voies. S'il existe une différence entre les bases sur un critère de segmentation, comme par exemple la taille minimum qu'un tronçon de route doit avoir pour être saisi en tant qu'entité indépendante, alors on peut avoir un conflit entre ces bases de données. Le dernier type de conflits est celui « de données ». En effet, les données sont saisies selon des processus de saisie différents d'une base de données à une autre. Ceci peut être engendré par des opérations de généralisation. (Abadie, 2012) préfère nommer ce type de conflit « conflit d'agrégation ». En plus des différences des niveaux de détail prévus pour les bases, des erreurs peuvent se produire pendant la saisie des géométries. De plus, certaines entités géographiques ont un caractère flou quand il s'agit de les délimiter géométriquement par un contour. C'est le cas pour les forêts ou les vallées. Tout cela engendre des représentations géométriques différentes entre les bases de données géographiques. Les différents conflits cités ici sont résumés dans la Figure 1.6.

Pour résumer, ces conflits sont principalement engendrés par les différences de niveaux de détail et de modélisation géométrique choisis pour les bases de données géographiques. Nous notons également le facteur humain dans la saisie des données géographiques. En effet, dans certains cas, la subjectivité ou la liberté d'interprétation laissée par certaines spécifications aux opérateurs de saisie, faisant confiance à leur expertise, peut engendrer des variations d'un opérateur à un autre (Abadie, 2012).

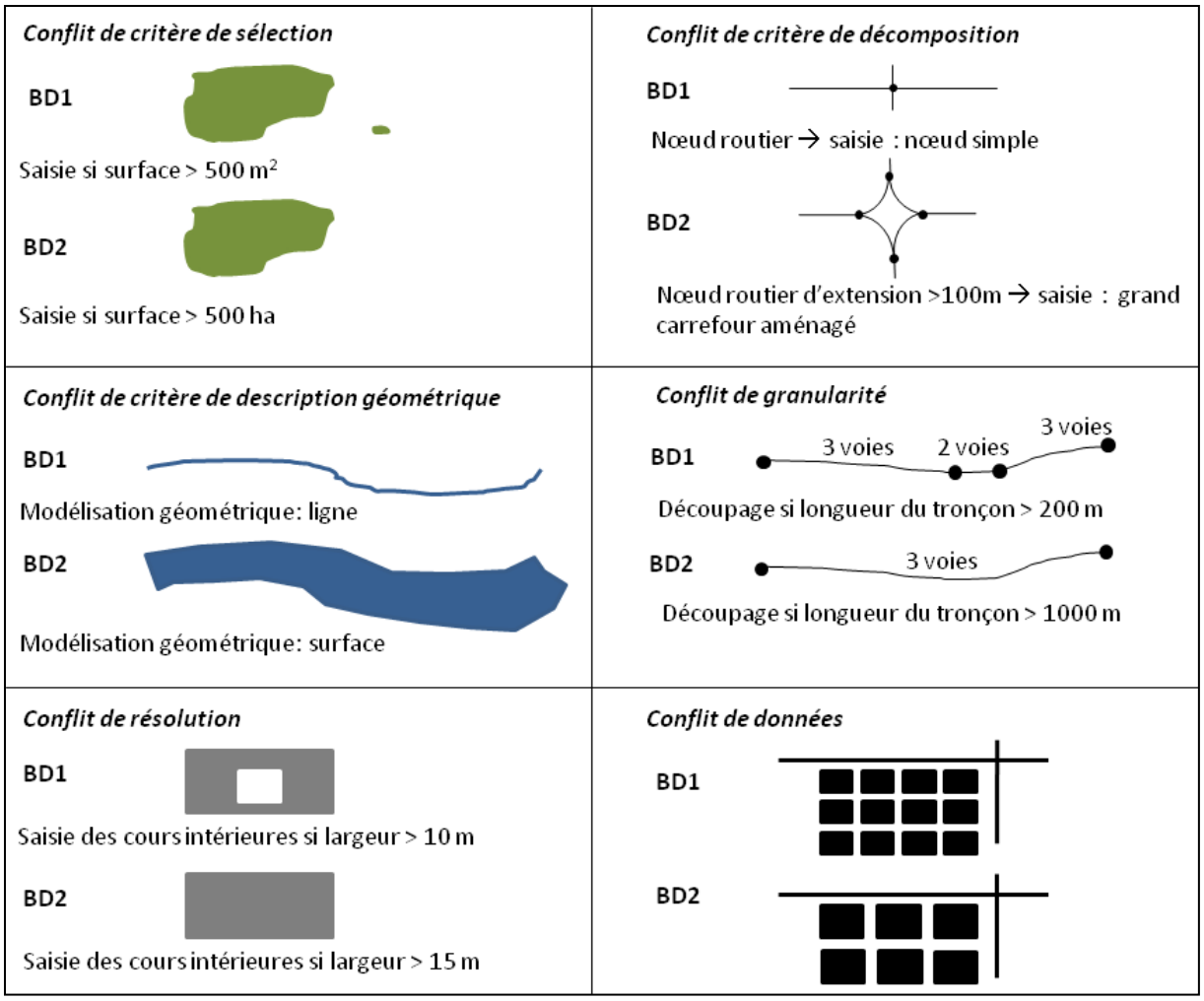


Figure 1.6 Exemples d'hétérogénéités géométriques des données géographiques (Abadie, 2012)

Les données géoréférencées du le Web ne sont pas toujours définies selon des spécifications ou des niveaux de détail bien précis. Les références spatiales utilisées demeurent généralement plus simples que celles que l'on peut trouver dans les bases de données géographiques, surtout quand il s'agit de sources de données d'origines a priori non spatiales, où les points sont le moyen le plus souvent utilisé pour géoréférencer les ressources. Cependant, les données géoréférencées du Web héritent des différents types d'hétérogénéités connus entre bases de données géographiques classiques.

Ainsi, on note toujours des différences de précision géométrique. La Figure 1.7 montre un exemple de cela : la ressource[54] DBpedia décrivant le Musée de l'Ordre de la Libération est localisée à plus de 100 mètres de l'Hôtel des Invalides qui l'héberge. Ce décalage important suggère qu'il s'agit plutôt d'une imprécision de localisation que d'une différence de choix de modélisation géométrique.

[54] http://fr.dbpedia.org/page/Musée_de_l'Ordre_de_la_Libération

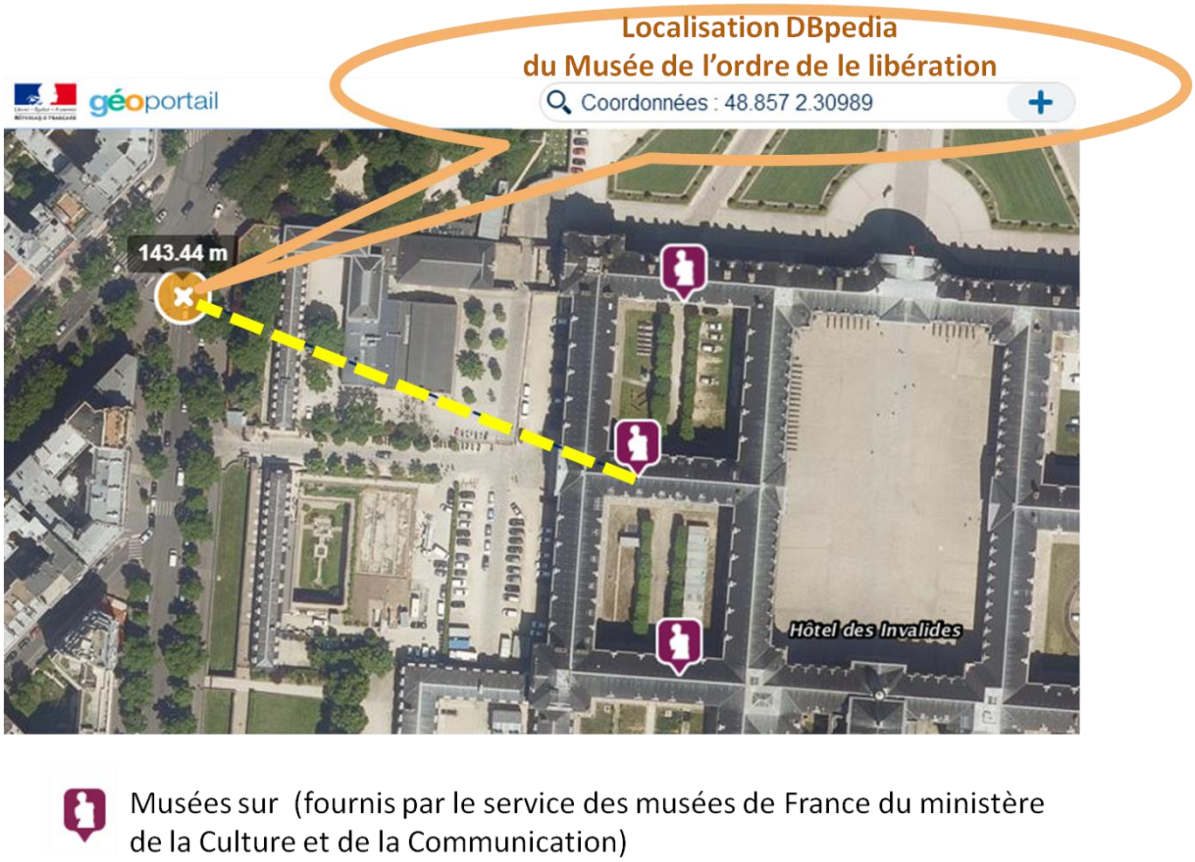


Figure 1.7 Exemple d'une imprécision de localisation des ressources géoréférencées du Web de données : la ressource DBpedia décrivant le Musée de l'Ordre de la Libération.

Les modélisations géométriques des mêmes entités géographiques peuvent être différentes d'une source à une autre : dans la Figure 1.8 on voit que dans la source Ordnance Survey (à gauche) la ressource décrivant Londres est représentée par une géométrie polygonale qui représente sa surface et sa frontière. Cependant dans DBpedia (à droite), la ressource décrivant Londres est localisée par un simple point situé sur la ville. En effet, la nature des deux sources joue un rôle dans cette différence de modélisation géométrique. Les données de l'Ordnance Survey sont des données de référence qui décrivent les unités administratives, dont la géométrie constitue une information centrale, ce qui nécessite une représentation géométrique détaillée. La source DBpedia n'étant pas de nature purement spatiale, la géométrie ne constitue pas une information centrale, ce qui explique l'utilisation d'une modélisation géométrique peu détaillée, sous forme de point, pour l'ensemble des entités géoréférencées de cette source.

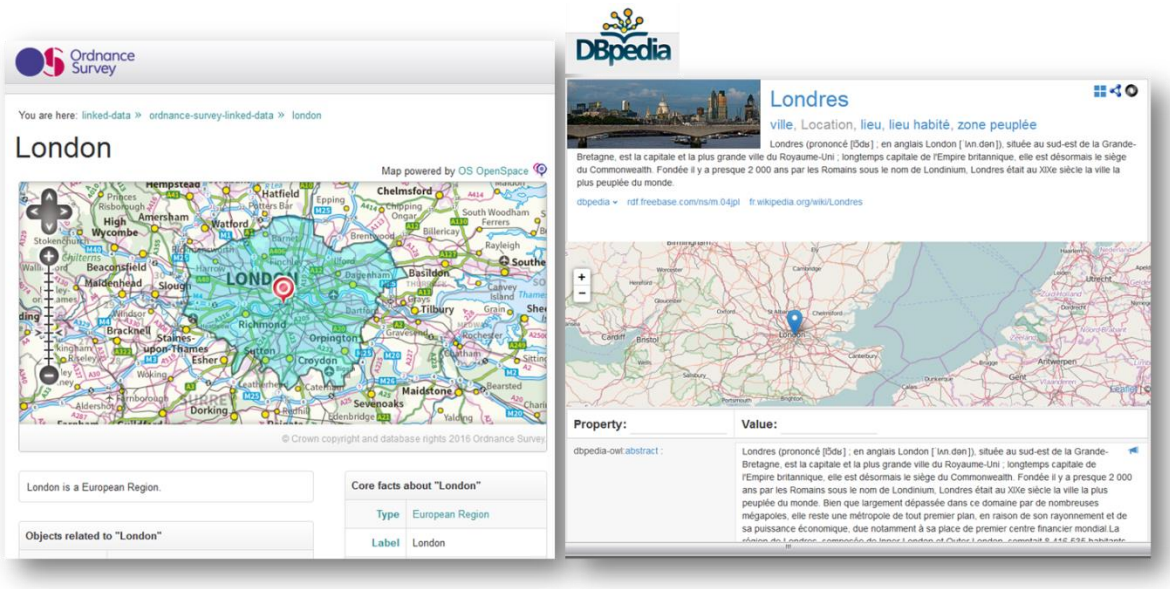


Figure 1.8 Exemple de différence de modélisations géométriques des références spatiales des ressources qui décrivent Londres entre DBpedia et l'Ordnance Survey.

En outre, les géométries peuvent être saisies avec des résolutions ou des granularités différentes d'une source à une autre comme le montre la Figure 1.9. Dans cet exemple on constate la différence de granularité dans la représentation des limites géographiques des pays entre la base de données de la nomenclature des unités administratives statistiques NUTS[55] et la base de données des régions administratives globales GADM[56]. Les géométries dans la source GADM (en vert) ont clairement un niveau de granularité plus élevé que celui des géométries dans la source NUTS (en rouge).

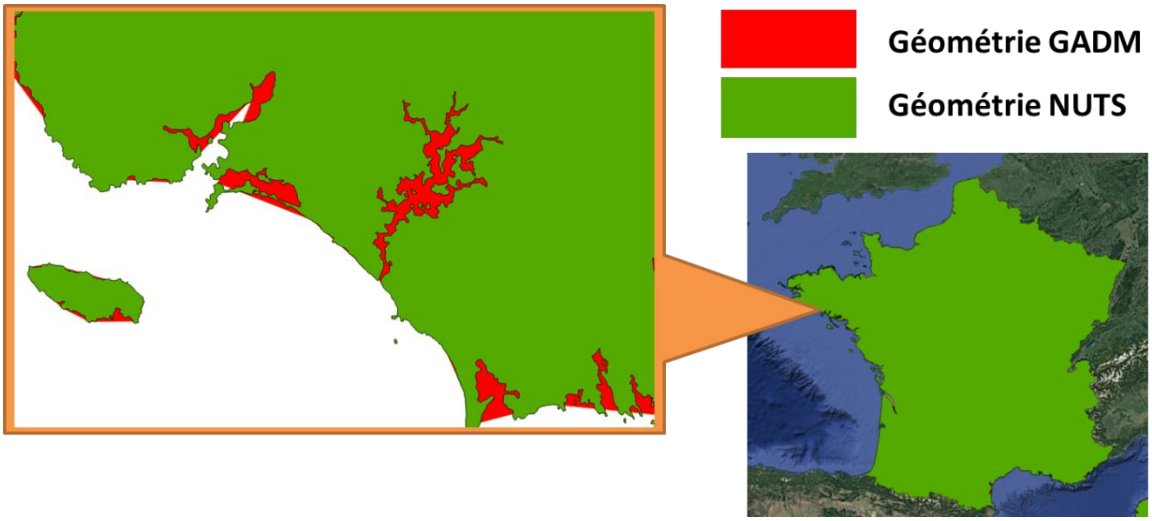


Figure 1.9 Exemple de différence de granularités géométriques entre les sources de données GDAM et NUTS

[55] http://ec.europa.eu/eurostat/web/nuts/
[56] http://gadm.org/

Enfin, le caractère vague de certains types d'entités géographiques engendre des difficultés de localisation et donc des représentations différentes entre les sources de données. La Figure 1.10 montre un exemple de représentation de col de montagne entre deux sources : les oronymes de la BD TOPO® et les points remarquables du relief de la BD CARTO®.

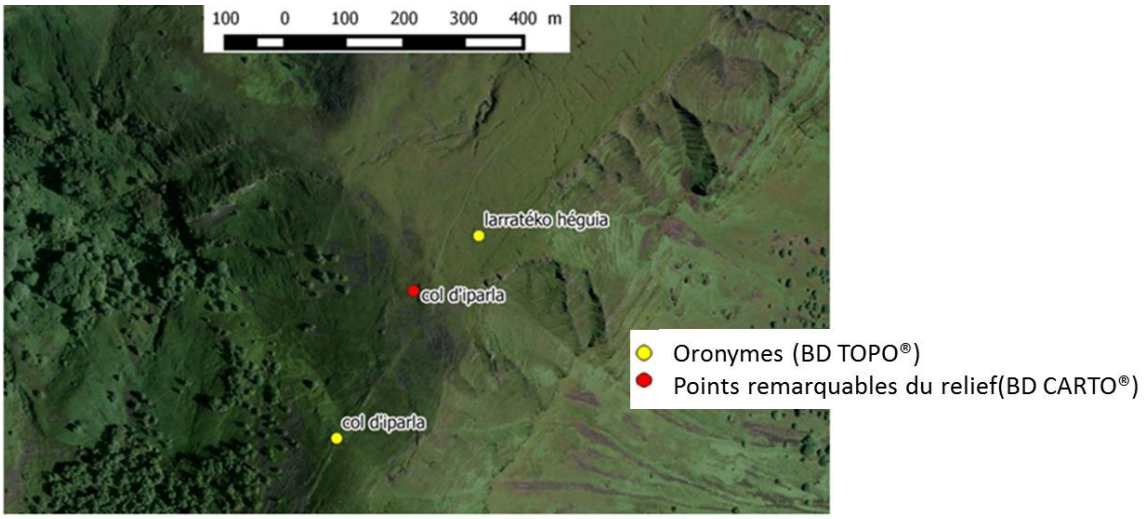


Figure 1.10 Exemple de différence de localisation d'un col de montagne entre la BD TOPO® et la BD CARTO® à cause du caractère vague de ce type d'entités géographiques.

Sur le Web, ces différences surviennent quand deux sources de données sont saisies selon des spécifications différentes, mais aussi en le cas d'absence ou de non-respect de ces spécifications. Prenons en exemple le chapitre francophone de DBpedia. Les localisations des ressources géoréférencées dans cette source sont majoritairement extraites de Wikipedia. Des recommandations pour la saisie de coordonnées géographiques pour les articles Wikipedia ont été proposées dans le cadre du projet « WikiProject Geographical coordinates »[57] . Celles-ci préconisent tout d'abord de recourir à des sources de confiance, comme les géoportails des agences cartographiques nationales, pour obtenir des coordonnées fiables. De plus, elles indiquent quel élément caractéristique de la forme de l'entité géographique décrite par chaque article doit être localisé en priorité, selon le type d'entité géographique considéré. Une localité devra donc plutôt être représentée par un point situé au centre de sa zone habitée et un bâtiment par un point situé à son entrée principale. Ces recommandations fournissent des règles d'arrondi des valeurs de coordonnées ainsi qu'une explication sur leur rajout dans un article. Le modèle prédéfini des *infobox* DBpedia prévoit la saisie de métadonnées pour chaque paire de coordonnées, comme l'échelle d'affichage la plus appropriée ou la source d'origine des coordonnées saisies. Cependant, rien ne garantit que ces règles et recommandations soient respectées. La Figure 1.7 en est un exemple concret, où le point n'est pas saisi à l'entrée du bâtiment comme le précisent les recommandations.

Un autre problème apparaît donc fréquemment dans les sources de données du Web, notamment celles ayant plusieurs origines, ou étant saisies d'une manière participative : c'est l'hétérogénéité interne. Ainsi, on voit dans l'exemple présenté dans la Figure 1.11 la représentation par des points de deux hôtels dans GeoNames. On constate que le premier est représenté par un point sur le bâtiment alors que le deuxième est représenté par un point sur la rue qui est en face de l'hôtel. On peut

[57] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geographical_coordinates

supposer qu'il s'agit de deux modélisations géométriques différentes ou deux précisions planimétriques différentes. Il s'agit en tout cas d'une hétérogénéité interne à la source de données.



Figure 1.11 Exemple d'hétérogénéité géométrique à l'intérieur d'une même source de données

Les données géographiques issues de plateformes de saisie collaborative sont souvent caractérisées par des niveaux de détail hétérogènes (Touya et Brando, 2013). La nécessité de faciliter le travail des contributeurs des plateformes collaboratives, tout en leurs permettant de gérer la cohérence du contenu produit, a été soulignée par (Brando, 2012). L'idée de proposer des contraintes d'intégrité pour la saisie collaborative proposée par (Brando, 2012) a été reprise par (Touya et Brando, 2013) dans la détection d'incohérences de niveau de détail dans une source de données collaborative. Cependant, on note l'absence de spécifications qui contraignent les contributeurs à respecter un seul niveau de détail. Les données liées géoréférencées dérivées de ces plateformes, telles que celles de LinkedGeoData, dérivées d'OpenStreetMap, héritent donc de cette hétérogénéité à l'intérieur d'une même source.

Ainsi, les approches d'interconnexion désireuses de mettre à profit les références spatiales pour la comparaison de ressources devront faire face aux mêmes types d'hétérogénéités que les algorithmes d'appariement de données géographique, mais devront en outre trouver des solutions pour résoudre ces hétérogénéités non plus entre sources de données, mais entre paires de ressources à comparer.

1.5 Géométries, liens et vocabulaires : visualiser les ressources géoréférencées du Web de données

La visualisation cartographique fournit l'un des moyens les plus pratiques pour l'exploration et la compréhension des données sur le Web, et donc l'amélioration de leur utilisabilité. Dans ce contexte, de nombreuses solutions de visualisation cartographique ont été proposées et varient des applications réservées à des jeux de données bien spécifiques aux applications génériques d'exploration de données. La plupart de ces applications utilisent les bibliothèques et technologies traditionnelles de la cartographie Web pour la visualisation de données publiées dans le modèle RDF.

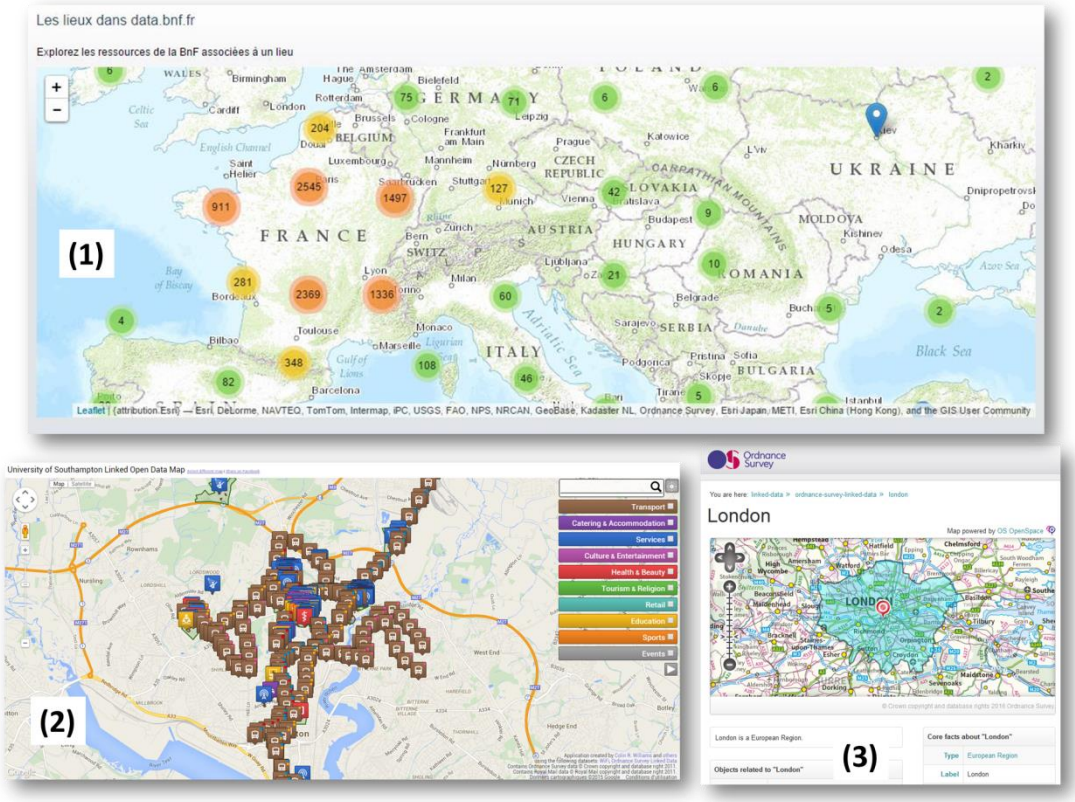


Figure 1.12 Exemple de solutions de visualisation cartographique propres à des sources de données géoréférencées du Web de données.

La Figure 1.12 présente des exemples de solutions de visualisation cartographique appliquées à des jeux de données spécifiques. Le premier exemple dans la Figure 1.12 représente la carte des ressources géoréférencées de la Bibliothèque Nationale de France (BNF)[58], tous thèmes confondus. Le deuxième exemple représente la localisation des données de la ville de Southampton[59] (stations de travail, points de recyclage, lieux de restauration, etc.). Le troisième exemple représente la carte localisant la région de Londres[60] sur le site de publication des données de l'Ordnance Survey. Ce genre d'applications est généralement fourni parallèlement à un jeu de données publié selon les bonnes pratiques du Web de données ou sur la page descriptive obtenue par le déréférencement de l'URI d'une ressource géoréférencée. Visualiser plusieurs ressources en même temps par affichage des points les points qui les localisent ou de symboles sur un fond cartographique, comme c'est le cas dans le deuxième exemple, est une pratique courante pour les solutions de visualisation de sources de données liées. D'autres applications, telles que les outils d'exploration cartographique de « Geo.LinkedData.es[61] » ou « LinkedGeoData.org » proposent ce type de visualisations cartographiques. Cependant, ce genre de cartes n'est souvent compréhensible qu'à des échelles de visualisation suffisamment grandes. La visualisation à petites échelles devient quasiment illisible en rasion de'une trop forte densité de points à afficher sur un espace restreint. Toutefois, nous notons que certaines solutions de visualisation, comme celle de la BNF, procèdent par agrégation des

[58] http://data.bnf.fr/
[59] http://opendatamap.ecs.soton.ac.uk/
[60] http://data.ordnancesurvey.co.uk/id/7000000000041428
[61] http://www.geo.linkeddata.es/browser.html

références spatiales des ressources pour assurer la lisibilité de la carte en gardant une cohérence entre la quantité d'information affichée et l'échelle de visualisation choisie. Dans le type d'application qui propose la visualisation de données géoréférencées d'un jeu de données, la Figure 1.13 montre l'exemple de l'application de visualisation spatio-temporelle des ressources de YAGO proposée par (Hoffart et al., 2013). Cette application permet de visualiser les localisations des ressources, sélectionnées par une requête, sur un fond cartographique en utilisant TimeMap[62].

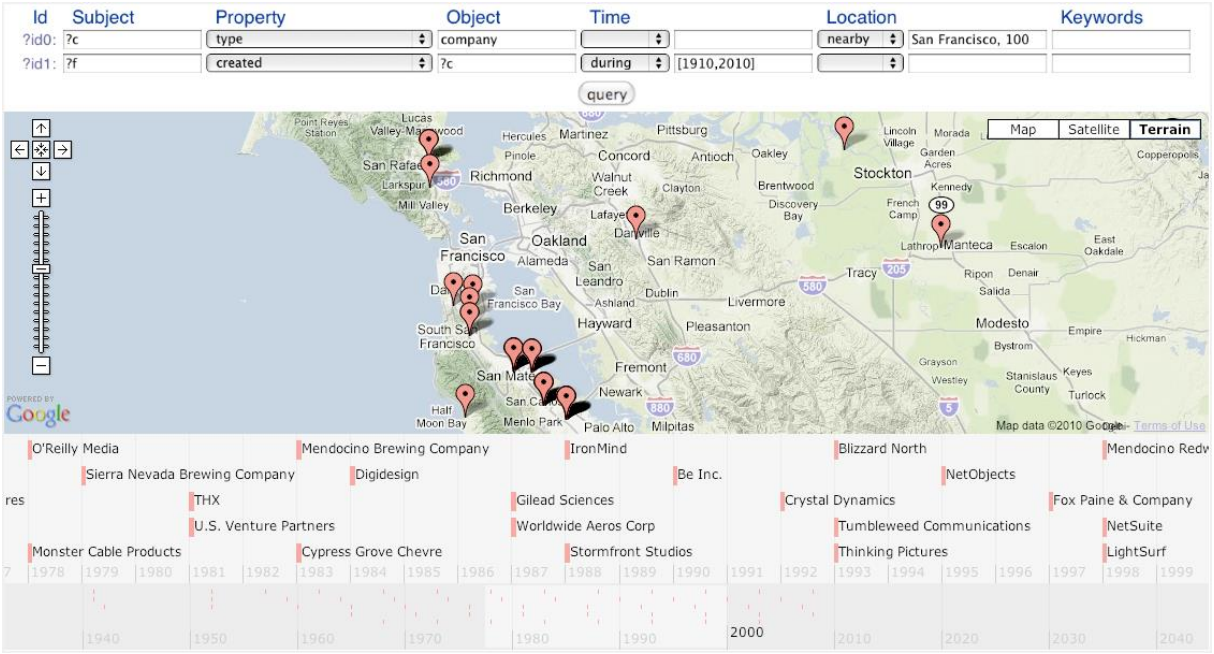


Figure 1.13 Visualisation des ressources YAGO qui représentent les entreprises fondées dans la région de la baie de San Francisco durant le dernier siècle (Hoffart et al., 2013).

D'autres applications plus génériques de visualisation cartographique de données géoréférencées sur le Web se présentent comme des moyens d'exploration et de navigation d'une source de données quelconque. La Figure 1.14 montre des exemples de ces solutions de visualisation. L'exemple de la partie supérieure de la Figure 1.14 est tiré de l'outil SemMap[63] qui permet l'exploration de jeux de données géographiques liées publiées sur le Web de données. En lui donnant en entrée l'URI du point d'accès d'une source de données géoréférencées, l'outil récupère l'arborescence des différentes classes (appelées facettes) contenues dans la source de données pour les afficher à l'utilisateur. Ce dernier permet, en sélectionnant une classe de ressources, de visualiser ses instances. Ainsi, l'outil propose une visualisation tabulaire des données en parallèle à la visualisation sur un fond cartographique. Afin de gérer la visualisation à différentes échelles, l'outil SemMap crée des blocs (*clusters*) dans les zones de la carte denses en ressources. Zoomer sur un bloc permet de visualiser chaque ressource indépendamment. Cependant, la carte reste illisible dans les zones de très forte densité de ressources géoréférencées. L'exemple de la partie inférieure de la Figure 1.14 est tiré de l'outil LodLive[64]. Cet outil est conçu pour permettre d'explorer visuellement des sources de données liées. Dans sa version actuelle, LodLive permet d'explorer une ressource à partir de son URI en affichant sa description RDF sous la forme d'un graphe. Ce graphe permet

[62] http://timemap.googlecode.com
[63] http://data.europa.eu/euodp/semmap/
[64] http://en.lodlive.it/

d'explorer les autres nœuds qui représentent les ressources liées à la ressource principale. Optionnellement, une visualisation sur un fond cartographique peut être réalisée si la ressource contient dans sa description des coordonnées géographiques. L'exemple montré dans la Figure 1.14 correspond à la ressource DBpedia qui représente la ville de Paris. La solution cartographique dans cet outil est conçue principalement pour la visualisation de la seule (ou des quelques) ressource(s) présente(s) dans le graphe en cours d'exploration.

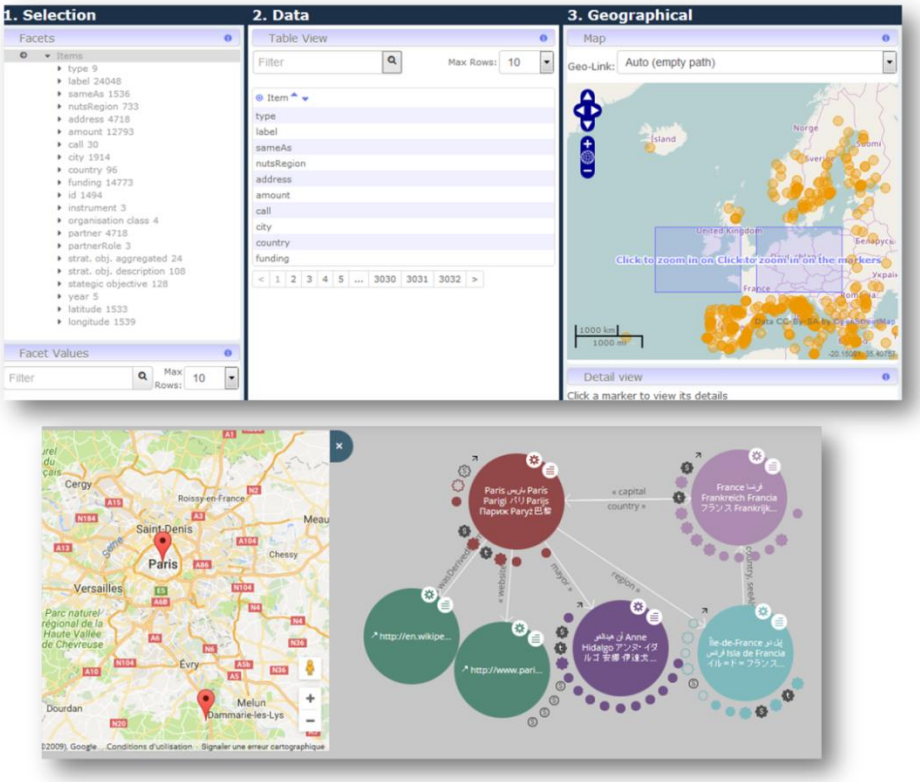


Figure 1.14 Exemple de solutions de visualisation cartographique qui permettent l'exploration des sources de données géoréférencées quelconques du Web de données

Les applications cartographiques liées à des sources spécifiques comme celles destinées à explorer différentes sources de données consistent habituellement à afficher des points sur un fond cartographique ou orthophotographique. Ceci est logique, du fait que la plupart des sources de données liées ne fournissent que des références spatiales ponctuelles. A l'exception de quelques exemples de solutions de visualisation cartographique de données liées, telles que celles mises au point par des fournisseurs d'autorité comme l'Ordnance Survey qui fournit une visualisation de géométries détaillées adaptée selon l'échelle, les solutions de visualisation cartographique de données liées ne sont que très peu adaptées à la visualisation multi-échelle. Ceci va à l'encontre du potentiel que la visualisation cartographique peut apporter en ce qui concerne l'exploration et la compréhension des données. Ce constat nous a poussés à réfléchir à des solutions cartographiques plus conviviales et plus adaptées à la visualisation multi-échelle. Nous suggérons de mieux exploiter ce que le Web de données permet d'offrir. Par exemple, l'interconnexion entre sources de données peut s'avérer très utile dans ce contexte pour combiner des données riches en informations thématique à des données géographiques riches en détail géométrique dans une visualisation

cartographique multi-échelle. Nous pouvons également tirer parti des vocabulaires qui décrivent les données afin dans les applications de visualisation cartographiques.

1.6 Objectifs de la thèse

Les liens qui associent les ressources des différentes sources de données constituent l'un des éléments les plus importants du Web de données, car ils favorisent un rapprochement d'informations, une complémentarité thématique et une navigabilité facile entre ces sources de données. Nous nous intéressons donc dans cette thèse au processus qui permet la création de ces liens : l'interconnexion. Ce processus s'appuie principalement sur la comparaison des descriptions de ressources pour identifier des similarités et décider de la création ou non de liens entre elles (Ferrara et al., 2013). Plusieurs problèmes se posent quand on vise à élaborer une approche d'interconnexion. L'hétérogénéité des données représente l'un des principaux défis de l'interconnexion. Dans le présent travail, nous nous intéressons plus principalement à l'un des éléments qui décrivent habituellement les ressources et qui sont utilisés dans la comparaison des ressources, à savoir les références spatiales. Nous avons vu que la nature ouverte du Web rend ses différentes sources de données beaucoup plus susceptibles de présenter des hétérogénéités géométriques internes que ne le sont les bases de données géographiques. Ceci rajoute un défi à celui des hétérogénéités géométriques entre sources de données, connu dans le domaine d'appariement de données géographiques.

L'objectif principal de ce travail est donc, en premier lieu, d'identifier et formaliser les éléments qui caractérisent les références spatiales et dont la différence d'une ressource à l'autre constitue une cause d'hétérogénéité. La formalisation de ces caractéristiques doit être ensuite mise en œuvre afin d'améliorer la fiabilité de la comparaison des références spatiales dans un processus d'interconnexion. La représentation formelle des connaissances qui décrivent les caractéristiques des références spatiales doit être conforme aux différents standards et normes de la représentation de l'information géographique ainsi qu'aux bonnes pratiques du Web de données. Les solutions proposées doivent tenir compte de l'existant en matière d'appariement de données géographiques et d'interconnexion des données sur le Web afin qu'elles soient utilisables pour le plus possible de cas d'application.

Nous nous intéressons en deuxième lieu à la visualisation cartographique des données liées géoréférencées. Les solutions de visualisation existantes, qui permettent d'afficher les données d'une source spécifique ou d'explorer des sources de données différentes, favorise la réutilisation des données, en permettant de découvrir visuellement leur contenu et de susciter chez les utilisateurs des idées sur les utilisations possibles des données. Cependant, ces solutions de visualisation cartographique se limitent dans leur majorité à l'affichage de points sur un fond cartographique. Or ceci peut poser des problèmes de lisibilité de la carte lorsque le niveau de détail des données affichées ne correspond pas à l'échelle de visualisation.

Notre second objectif est donc de proposer des solutions de visualisation cartographique qui restent lisibles à différentes échelles. Il s'agira donc de mettre à profit l'expressivité du modèle du Web de données et l'état de l'art important en généralisation de l'information géographique pour offrir des solutions conviviales de découverte cartographique de données liées.

2 ÉTAT DE L'ART : APPARIEMENT DE DONNÉES GÉOGRAPHIQUES ET INTERCONNEXION DES DONNÉES SUR LE WEB

Cette partie dresse un état de l'art sur l'appariement de données géographiques ainsi que l'interconnexion des données du Web. Ces deux domaines partagent le même objectif de mise en correspondances d'objets géographiques ou de ressources homologues. Dans la plupart des cas, cette mise en correspondance est réalisée par comparaison élémentaire des valeurs des attributs des objets ou des propriétés des ressources. La comparaison des attributs ou des propriétés se fait en calculant l'écart entre leurs valeurs grâce à des mesures de distances, ou en calculant la proximité entre leurs valeurs en utilisant des mesures de similarité. Une mesure de distance (resp. similarité) est une fonction qui associe une paire de valeurs à une valeur numérique qui quantifie l'écart (resp. la proximité) entre elles. Les notions de mesures de distance et de similarité sont fortement liées puisque l'une est souvent définie comme étant l'inverse de l'autre. Les notions de métriques de distance ou de similarité et la relation qui existe entre elles sont définies par (Shihyen Chen et al, 2009) et (Euzenat et shvaiko, 2007) d'une manière plus formelle. Les définitions des métriques ont donc un sens mathématique plus spécifique. Une métrique de distance doit respecter, pour des valeurs x, y et z, des conditions de non-négativité ($d(x,y) \geq 0$), de symétrie ($d(x,y) = d(y,x)$), d'inégalité triangulaire ($d(x,z) \leq d(x,y) + d(y,z)$) et d'identité des indiscernables ($d(x,y) = 0\ ssi\ x = y$). Dans cette thèse, nous utilisons la notion de mesure, car nous pensons que les mesures proposées dans la pratique pour calculer l'écart ou la proximité de valeurs ne respectent pas forcément toutes les conditions de la définition formelle de métrique (ex. la condition d'identité des indiscernables n'est pas toujours respectée). Nous utiliserons alternativement les termes distance ou similarité selon l'interprétation du sens de la mesure concernée. Les mesures de distance ou de similarité sont dites normalisées si elles fournissent uniquement des valeurs dans l'intervalle [0, 1]. Une distance normalisée peut être facilement transformée en similarité par le calcul suivant : $similarité = 1 - distance$. Dans une tâche de mise en correspondance, deux vérifications principales sont nécessaires afin d'associer deux objets (ou ressources) *a* et *b*: la distance (resp. une similarité) entre *a* et *b* est inférieure (resp. supérieure) à un seuil choisi, tout en étant minimale (resp. maximale) pour *a* et *b*. La notion de seuil de distance ou de similarité est donc très importante pour la tâche de mise en correspondance.

Après la présentation de quelques généralités sur les approches proposées dans les deux domaines, nous détaillons dans cette partie, pour chaque domaine, les différentes propositions de l'état de l'art qui sont susceptibles de répondre, même partiellement, aux objectifs de ce travail de thèse. Nous conclurons par une synthèse des avantages et limites des approches existantes pour résoudre l'hétérogénéité géométrique des données géoréférencées publiées sur le Web.

2.1 Appariement de données géographiques

Dans le domaine des sciences de l'information géographique, l'appariement est le processus qui vise à mettre en correspondance des objets géographiques issus de bases de données hétérogènes qui représentent le même phénomène du monde réel (devogele, 1997 ; Walter et Fritsch, 1999). On cherche donc à trouver des similarités dans la représentation des entités du monde réel entre bases de données géographiques hétérogènes. La comparaison entre objets de base de données géographique s'effectue principalement sur trois critères : la similarité des géométries des objets, la similarité des valeurs des attributs non géométriques des objets, et la similarité des relations

topologiques (les voisinages) des objets. Les critères non géométriques concernent les attributs quantitatifs (*ex.* la population, la densité) ou qualitatifs (*ex.* le label, la nature) qui décrivent les objets géographiques. La comparaison des géométries se fait sur plusieurs caractéristiques comme la localisation, la forme, l'orientation ou encore la taille. Les géométries étant l'élément central dans les bases de données géographiques, de nombreux travaux se sont intéressés aux meilleurs moyens de les comparer et de rechercher des similitudes entre celles-ci. Cette comparaison s'effectue en utilisant des mesures de distance adaptées aux types des géométries comparées. De nombreuses mesures de distances ont été proposées dans la littérature et nous les discutons dans la suite de cette section. Nous présenterons ensuite les différentes approches d'appariement de données géographiques selon les stratégies adoptées pour résoudre les problèmes d'hétérogénéité géométrique.

L'appariement de données géographiques est très important dans le domaine des données géographiques, car il répond à plusieurs besoins (devogele, 1997; olteanu, 2008). L'une des utilisations les plus importantes de l'appariement reste l'intégration de bases de données géographiques. L'intégration de données cherche à combiner les données de différentes sources, et fournir à l'utilisateur une vue unifiée de ces données (Lenzerini, 2002). Dans le cas de bases de données géographiques, la définition reste la même, même si les techniques utilisées sont propres à ce domaine. Un processus d'intégration de données est constitué de trois étapes : la préparation des données, l'appariement des schémas et des données puis l'intégration finale des schémas et des données (Devogèle et al., 1998). Le but de l'intégration peut être l'obtention d'une seule base de données finale avec un schéma unifié et des données non redondantes. Il peut être aussi de garder le schéma de l'une des bases à intégrer et de transformer toutes les données des autres, sans redondance, dans ce schéma. Le but final de l'intégration peut être la création d'un seul schéma fédéré auquel se relient les schémas des bases intégrées. Dans ce cas un système fédéré permet d'accéder à toutes les données selon ce schéma fédéré. L'appariement peut être utilisé à des fins de contrôle de qualité (interne) des données *c.-à-d.* la vérification de la conformité des données aux spécifications de leur producteur. L'appariement sert dans ce cas à associer chaque objet à contrôler à (aux) l'objet (objets) correspondant(s) dans le jeu de données de référence. Ensuite, les objets à contrôler peuvent être comparés à leurs objets correspondants dans le jeu de données de référence pour évaluer leur qualité. L'appariement peut être aussi un moyen de propagation de mises à jour d'une base de données géographique de références aux bases de données utilisateur dérivées de cette base. Apparier les objets des jeux de données utilisateurs aux objets correspondants dans la base de données de référence permet de propager les mises à jour de cette dernière seulement aux objets concernés dans les bases dérivées. Apparier un jeu de données à une base de données de référence peut être utilisé également à des fins de recalage de géométries. Le recalage est un processus qui vise à joindre les géométries de deux jeux de données géographiques représentant une même entité du monde réel de sorte à ce qu'elles se superposent (Méneroux et Brasebin, 2015). Ce processus peut s'appliquer sur les données Raster comme sur les données Vecteur. Dans le cas des données Vecteur, l'appariement sert à trouver des points homologues entre le jeu à recaler et la base de données de référence. Ces points sont utilisés dans le calcul d'une transformée de recalage qui est appliquée ensuite sur l'ensemble des géométries du jeu de données à recaler.

2.1.1 Des mesures de distance pour comparer les objets géographiques

L'utilisation des mesures de distance et de similarité entre les valeurs d'attributs a fait l'objet de nombreux travaux dans le domaine des bases de données ainsi que d'autres disciplines (*ex.* la bio-informatique). On dispose donc d'une grande variété de mesures adaptées aux types des valeurs comparées et aux finalités applicatives. Dans cette section, nous mettons l'accent sur la comparaison des géométries qui constitue un défi spécifique à l'appariement des données géographiques. De nombreuses mesures de distances ont été proposées pour comparer deux géométries selon leur type de primitive géométrique. Dans certaines approches d'appariement, les géométries doivent subir des transformations pour pouvoir les comparer. Nous présentons quelques opérateurs de transformation de géométries, ainsi que les principales mesures de comparaison adaptées à chaque type de primitive géométrique.

Opérateurs de transformation de géométries

La transformation des géométries est souvent nécessaire afin de permettre leur comparaison. Une première transformation, souvent nécessaire, est celle qui vise à convertir des géométries dans un même système de coordonnées de référence (SCR). En effet, disposer de géométries définies dans un même système de coordonnées est primordial pour permettre leur comparaison. En outre, les mesures de distance sont souvent adaptées à un seul type de SCR : *ex.* la distance euclidienne qu'on verra par la suite n'est adaptée qu'à des coordonnées représentées dans système de coordonnées de référence projeté.

L'une des transformations les plus utilisées dans le cas des réseaux (*ex.* réseaux routiers) consiste à extraire un graphe topologique à partir des géométries (*ex.* (Walter et Fritch, 1999 ; Mustière et Devogele, 2008)). Dans ce cas, les géométries sont transformées sous la forme d'un graphe où les lignes sont considérées comme des arêtes et leurs limites comme des sommets. Ceci permet la prise en compte des relations topologiques entre les géométries d'une même base lors de leur comparaison avec les géométries d'une autre base, dont les relations topologiques sont également connues.

Un autre type de transformations qui peuvent être utiles pour comparer les géométries sont les opérateurs de généralisation. Ceux-ci sont utiles dans le cas où une cohérence du niveau de détail des géométries est nécessaire pour pouvoir les comparer. L'harmonisation du niveau de détail des géométries peut s'effectuer en utilisant des opérations de généralisation (Touya et Brando, 2013). L'approche proposée par (Yang et al., 2014) par exemple, pour l'appariement des réseaux routiers de niveaux de détail différents, utilise une opération de simplification des intersections complexes et des autoroutes dont chaque voie est représentée par une polyligne distincte en une représentation abstraite en simples points et lignes. Selon les auteurs, cette représentation abstraite constitue des « patrons » qui restent invariants pour les réseaux routiers ayant des niveaux de détails différents. La Figure 2.1 représente quelques exemples des nombreux opérateurs de généralisation décrits par la classification de (Mustière, 2001) : Il s'agit d'opérateurs de simplification, de caricature ou d'harmonisation.

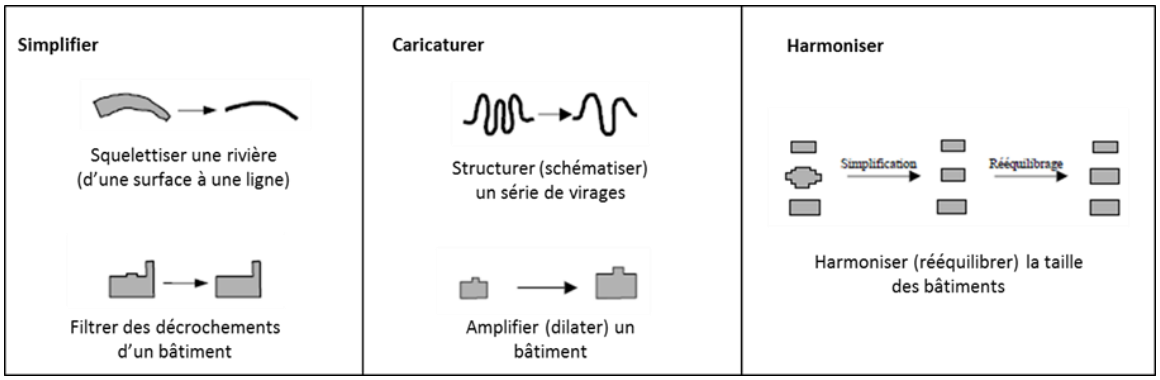


Figure 2.1 Exemples d'opérateurs de généralisation issus de la classification de (Mustière, 2001)

Mesures de distance entre primitives ponctuelles

La distance euclidienne est la mesure basique utilisée pour calculer l'écart minimum entre deux points sur un plan. Entre deux points P_1 et P_2 de coordonnées respectives (x_1, y_1) et (x_2, y_2), la distance euclidienne est définie par l'équation 2.1

$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad 2.1$$

La distance euclidienne est adaptée seulement aux coordonnées représentées dans un SCR projeté. Or, les coordonnées d'un point peuvent être aussi exprimées dans un système de coordonnées géographiques. Elles représentent alors sa position sur un ellipsoïde terrestre (*ex.* WGS84) a des angles de longitude et de latitude. Il est conseillé dans ce cas de commencer par une projection des coordonnées de longitude et de latitude dans SCR projeté, ou d'utiliser une mesure de distance sur un ellipsoïde, telle que celle proposée par (Vincenty, 1975; Deakin and Hunter, 2007).

Mesures de distance entre primitives linéaires

La comparaison de polylignes est une tâche plus ardue que la comparaison de points. D'autres caractéristiques que la localisation, comme la forme et la longueur, peuvent être prises en compte pour chercher des similarités entre les géométries dans ce cas. Nous mentionnons ici les mesures de distance principales proposées dans la littérature.

La distance de *Hausdorff* est l'une des distances entre polylignes les plus connues. Entre deux géométries G_1 et G_2, cette distance est basée sur les maximums des écarts minimaux entre chaque point de G_1 et la géométrie G_2, et chaque point de G_2 et la géométrie G_1, comme expliqué dans la Figure 2.2 et l'équation 2.2.

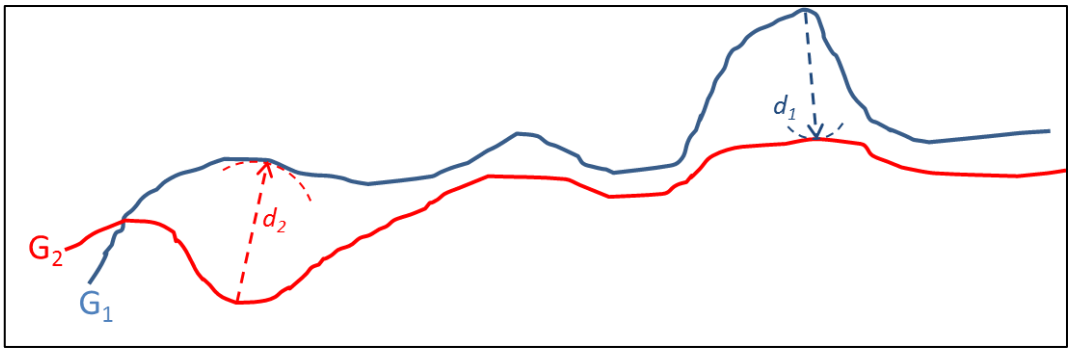


Figure 2.2 Exemple explicatif de la distance de *Hausdorff*

$$d_{Haus} = \max(d_1, d_2)$$
$$d_1 = \max_{p_1 \in G_1} \left(\min_{p_2 \in G_2} \left(d_E(p_1, p_2) \right) \right) \quad 2.2$$
$$d_2 = \max_{p_2 \in G_2} \left(\min_{p_1 \in G_1} \left(d_E(p_2, p_1) \right) \right)$$

La distance *moyenne* proposée par (McMaster, 1986} est une autre mesure de distance entre deux polylignes. Elle prend en compte leurs longueurs ainsi que la surface qui les sépare pour quantifier l'écart entre elles. Entre deux géométries G_1 et G_2 de longueurs l_1 et l_2, elle est calculée selon l'équation 2.3 comme expliqué dans la Figure 2.3

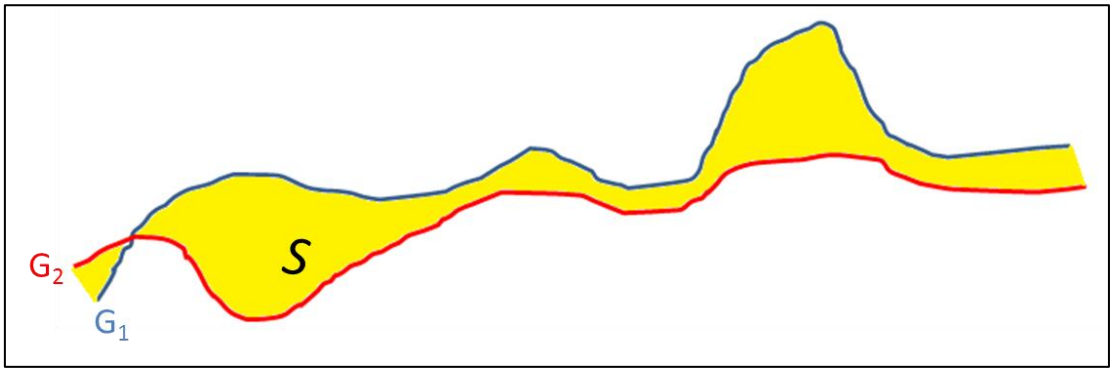


Figure 2.3 Exemple explicatif de la distance moyenne

$$d_{moy} = \frac{S}{\frac{l_1 + l_2}{2}} \quad 2.3$$

Cette distance est plus facile à calculer et a été conçue principalement pour calculer la distance dans le cas où G_2 a été générée par un processus de généralisation à partir de G_1.

La distance de *Hausdorff* présente des limites en cas de polylignes sinueuses : elle peut être très petite pour des polylignes qui ne sont pas similaires (*ex.* Figure 2.4). La distance dite de *Fréchet*, proposé par (Alt et Godau, 1995}, est mieux adaptée à ces. Elle est calculée selon l'équation 2.4. Les deux polylignes G_1 et G_2 (composées respectivement de N et M segments) sont considérées comme deux fonctions $f_1: [0, N] \rightarrow V_1$ et $f_2: [0, M] \rightarrow V_2$ tel que V_1 et V_2 sont des espaces vectoriels. $f_1(i)$ et $f_2(j)$ représente les sommets des segments de G_1 et G_2.

$$d_{Frech} = \min_{\alpha,\beta}\left\{\max_{t}[d(f_1(\alpha(t)), f_2(\beta(t)))]\right\} \quad 2.4$$

d représente une distance (euclidienne par exemple). α(t) et β(t) sont des fonctions continues et croissantes avec le temps, avec α(0)=0, β(0)=0, α(1)=N et β(1)=M.

Pour mieux expliquer l'idée de la distance de *Fréchet* , (Devogele, 1997) l'illustre par l'exemple d'un chien et son maître qui se déplacent sur deux chemins : « Ils avancent ou s'arrêtent à volonté, indépendamment l'un de l'autre, mais ils ne peuvent pas revenir sur leurs pas. La distance de Fréchet entre ces deux chemins est la longueur minimale de la laisse qui permet de réaliser une progression de concert satisfaisant ces conditions. »

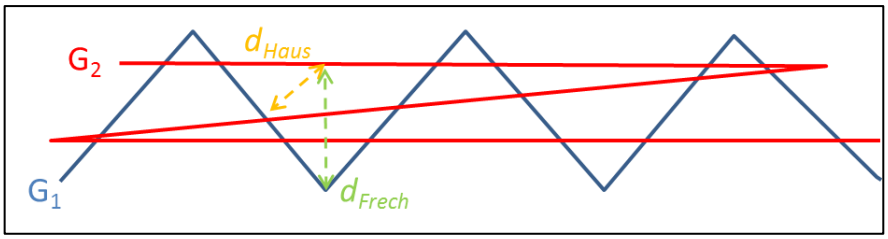


Figure 2.4 Exemple de cas où la distance de *Hausdorff* n'est pas adaptée pour comparer les géométries linéaires. On lui préfère alors la distance de *Fréchet.*

Un autre exemple distance entre polylignes est l'utilisation d'une bande epsilon (une zone tampon autour d'une polyligne) afin de trouver les géométries homologues dans des travaux comme (Sui et al., 2004).

Mesures de distance entre primitives surfaciques

(Vauglin, 1997} propose une mesure de similarité surfacique pour comparer des polygones, basée sur le rapport entre l'aire de l'intersection de deux polygones et l'aire de leur union. Pour deux polygones P_1 et P_2, elle est calculée ainsi :

$$d_{surf} = 1 - \frac{Aire(P_1 \cap P_2)}{Aire(P_1 \cup P_2)} \quad 2.5$$

Cette distance ne prend en compte que la superficie partagée entre les polygones, mais ne prend pas en compte d'autres caractéristiques comme la forme et l'orientation. Pour cela, on peut utiliser d'autres distances comme la distance entre signatures polygonales ou la différence entre fonctions angulaires. La fonction à distance radiale (ou signature polygonale) (Cohen et Guibas, 1997) trace la distance entre les points constituant un polygone et son centre de gravité. La fonction angulaire (Arkin et al., 1991) quant à elle trace les mesures des angles formés par les segments d'un polygone. Les fonctions à distance radiale, comme les fonctions angulaires, calculées pour deux polygones, doivent être mises en phase pour pouvoir les comparer.

Comparaison topologique

Les relations spatiales qui existent entre les objets d'une base de données géographiques ne sont le plus souvent pas représentées d'une manière explicite dans celle-ci. L'importance de représenter et traiter les relations spatiales a néanmoins été soulignée par des travaux comme (Papadias et Theodoridis 1997), (Mackaness et al. ,2013) et (Jaara, 2015). La modélisation des relations

topologiques a suscité l'intérêt de plusieurs travaux comme (Egenhofer et Franzosa,1991 ; Egenhofer et Herring, 1991 ; Egenhofer et AlTaha, 1992). (Egenhofer, 1990) propose un modèle de représentation des relations topologiques. Ce modèle explicite, via une matrice 3x3, les relations que peuvent avoir les intérieurs, les frontières et les extérieurs de deux géométries comparées. Entre différentes primitives géométriques, ce modèle permet d'identifier des relations comme « contient », « touche », « intersecté », « disjoint de», etc.

Les relations topologiques sont utilisées également pour comparer les voisinages des objets géographiques. C'est le cas des approches d'appariement de réseaux où ces derniers sont modélisés sous la forme de graphe topologique. Dans des approches comme celles de (Walter et Fritch, 1999 ; Mustière et Devogele, 2008), l'appariement des arcs entre deux graphes s'appuie sur l'appariement des sommets qui sont en relation avec ces arcs. L'approche d'appariement proposée par(samal et al. ,2004) utilise également la comparaison des voisinages des objets comme l'un des critères de comparaison des objets eux-mêmes.

2.1.2 Stratégies d'appariement de données géographiques

L'état de l'art sur l'appariement des données géographiques comprend un grand éventail d'approches qui diffèrent les unes des autres par le besoin auquel elles répondent. Il n'existe pas d'approche générique qui répond à tous les besoins. Chaque approche se distingue par la (les) hétérogénéité(s) à laquelle(auxquelles) elle fait face, les critères de comparaison utilisés, les types de primitives géométriques des données, leur mode de représentation, les performances en matière d'utilisation des ressources et du temps, etc. L'appariement peut paraître facile si on considère que deux bases de données distinctes sont conçues et saisies d'une manière similaire, *c.-à-d.* selon le même niveau de détail et les mêmes choix de modélisation géométrique. Or, dans les faits, les bases de données géographiques qui représentent une même réalité peuvent s'avérer très hétérogènes (*cf.* section 1.4.2). Le concepteur d'un algorithme d'appariement est souvent un expert des données qui comprend leurs hétérogénéités. L'algorithme proposé par ce concepteur s'appuie sur des hypothèses qui contraignent son utilisation à un contexte précis. Il est défini pour remédier à des hétérogénéités bien définies, comme proposer une approche pour apparier des données géoréférencées par des points ayant des précisions planimétriques différentes. Utiliser une approche d'appariement nécessite donc des connaissances sur les caractéristiques des jeux de données, et donc sur les hétérogénéités qui découlent des différences entre ces caractéristiques, pour vérifier leur adéquation aux hypothèses sur lesquelles repose l'approche envisagée. Comme nous le verrons dans la suite, la prise en compte des connaissances sur les données se manifeste d'une manière différente d'une approche d'appariement à une autre : certaines approches proposent des algorithmes ad hoc qui interprètent d'une manière implicite les connaissances du concepteur sur les données, alors que d'autres approches tendent vers une représentation explicite formelle de ces connaissances afin de les exploiter dans l'appariement.

Algorithme d'appariement avec des a priori sur les données géographiques en entrée

Connaître les différences des caractéristiques des données géographiques, notamment des les caractéristiques des géométries, est indispensable pour définir la stratégie d'un algorithme d'appariement. Dans de nombreuses approches d'appariement, la stratégie et les mesures mises en

œuvre témoignent de l'expertise du concepteur de l'approche sur les données à traiter et les difficultés particulières qu'elles posent. Bien que les connaissances du concepteur sur les données ne soient pas explicitement exprimées dans ce cas, elles demeurent ancrées dans les algorithmes proposés.

L'approche proposée par (Walter et Fritch, 1999) en est un exemple. Cette approche proposée pour l'appariement de réseaux (comme les réseaux routiers) suppose que les géométries des deux jeux de données à apparier ont la même granularité, mais des précisions différentes. Pour remédier à ce problème, plusieurs techniques sont utilisées dans cette approche. Au début, les données sont transformées dans une représentation topologique sous forme de graphes (arcs et sommets). Des paires de points de contrôle doivent ensuite être sélectionnées manuellement. Ces paires de points sont utilisées pour faire face au problème de différence de précisions géométriques tout au long de cette approche. Elles sont utilisées pour effectuer un recalage pour corriger le décalage géométrique entre les deux jeux de données. Les écarts entre ces paires de points sont également utilisés pour calculer un seuil de distance qui sert à filtrer les couples d'appariement improbables. Ces couples d'appariement sont sélectionnés en cherchant l'inclusion des géométries du deuxième jeu de données dans des zones tampons grandissantes autour des géométries du premier jeu de données. La dernière étape consiste à choisir des couples optimaux par adaptation de la théorie de l'information. Ceci est réalisé en calculant, pour chaque paire de candidats, la somme des informations mutuelles calculées pour chaque critère géométrique (la localisation, la forme, la longueur et l'angle). Pour cela, chaque critère est associé à une probabilité estimée à partir des couples déjà appariés. Les paires de candidats ayant des sommes d'information maximales sont considérées comme homologues. À travers ces différentes techniques, cette approche permet de réduire principalement l'effet de la différence de précision planimétrique entre les géométries des deux jeux de données dans le processus d'appariement. Elle reste cependant incapable de gérer les différences de granularité ou de modélisation géométrique. Elle est également fortement dépendante des paires de points appariés manuellement.

Dans (Mustière et Devogele, 2008), l'approche proposée prend en compte la différence des niveaux de granularité des géométries entre un réseau et un autre plus détaillé. L'hypothèse de base de cette approche est que le premier réseau (moins détaillé) est complètement couvert par le deuxième réseau (plus détaillé). La prise en compte des différences de granularité des géométries entre les deux jeux de données transparaît dans la proposition d'un algorithme à plusieurs étapes. Une première étape consiste à trouver, pour chaque nœud du réseau le moins détaillé, un ensemble de nœuds candidats dans le réseau le plus détaillé se trouvant à une distance euclidienne inférieure à un seuil fixé. La même opération est effectuée ensuite entre les arcs des deux réseaux en utilisant la distance de Hausdorff. Le choix des appariements des nœuds est décidé en cherchant pour chaque nœud les cohérences entre les deux pré-appariements précédents : pour chaque nœud N_1 dans le graphe le moins détaillé on vérifie si les nœuds N_2 du graphe le plus détaillé, pré-appariés à N_1, sont des candidats « complets », « incomplets » ou « impossibles ». N_2 est considéré complet si on peut trouver un lien de pré-appariement entre tous les arcs connectés à N_1 et des arcs connectés à N_2. N_2 est considéré incomplet si on peut trouver un lien de pré-appariement entre certains arcs connectés à N_1 et certains arcs connectés à N_2. N_2 est considéré impossible sinon. La décision de la validité et de la cardinalité des liens avec entre le nœud N_1 et les nœuds candidats N_2 dépend de cette classification. Par exemple si pour N_1, il existe plusieurs nœuds N_2 qui sont complets, un lien de cardinalité 1:1 est créé avec le plus proche de ces nœuds. Si N_1 a plusieurs candidats incomplets, on

peut vérifier si une relation de cardinalité 1:n peut être créée entre N_1 et tout (ou une partie de) l'ensemble de ses candidats incomplets. La validation des appariements finaux des arcs est finalement guidée par les appariements des nœuds de l'étape précédente. Bien que la différence de granularité des données implicitement prise en compte par la possibilité d'apparier chaque nœud (respectivement arc) du réseau moins détaillé avec plusieurs nœuds (respectivement arc) du réseau plus détaillé, et par le choix du seuil de pré-appariement (qui nécessite tous les deux une connaissance préalable de données), cette approche reste restreinte au cas d'inclusion d'un réseau dans un autre. Plus important encore, cette approche suggère une homogénéité interne des jeux de données, et ne répond donc pas au défi de l'hétérogénéité interne aux jeux de données.

(Volz, 2006 propose une approche qui s'applique à des réseaux représentés également selon une organisation topologique. L'approche commence par un recalage d'un réseau sur un autre. Afin de favoriser des relations de cardinalité 1:1, l'algorithme commence par homogénéiser la granularité des deux jeux de données en rajoutant des nœuds à chaque réseau par projection des nœuds de l'un des deux réseaux sur les arcs proches de l'autre réseau (c.-à-d. situés à une distance inférieure à un seuil fixé). Les nœuds les plus proches (dans les deux sens) sont ensuite appariés avec des relations de cardinalité 1:1. Les relations entre ces nœuds sont ensuite employées pour détecter des relations de cardinalité 1:1 entre les arcs en relation avec ces nœuds. L'étape suivante consiste à chercher des relations de cardinalité 1:2 pour les nœuds qui n'ont pas été appariés dans la première phase. La même chose est ensuite appliquée aux arcs.

(Li et Goodchild, 2011) propose une approche d'appariement de données qui utilise une variante de la distance de Hausdorff qui prend en compte les angles entre les lignes pour comparer les géométries linéaires, dans une approche multicritère utilisant également les noms des objets géographiques comparés via la distance de *Hamming*. L'idée principale de l'approche est l'utilisation d'une fonction objective globale qui, en la maximisant, garantit un appariement optimal.

(Yang et al., 2014) propose également une approche pour l'appariement des réseaux routiers de niveaux de détail différents. Cette approche utilise une opération de simplification pour réduire l'effet de la différence des niveaux de détail. Comme décrit précédemment (*cf.* section 2.1.1), on commence par simplifier des intersections routières complexes et des autoroutes, dont les différentes voies sont représentées par différentes polylignes, en proposant une représentation abstraite en simples points et lignes, que les auteurs appellent des « patrons ». Ces patrons restent invariants pour les réseaux routiers quel que soit leur niveau de détail d'origine, et constituent donc un moyen de comparer des réseaux routiers ayant des niveaux de détail différents. L'idée de l'utilisation des patrons de simplification semble prometteuse pour pallier l'effet de la différence des niveaux de détail, mais reste limitée dans ce travail aux cas des intersections des routes et des autoroutes à plusieurs voies.

L'approche proposée par (Costes, 2014) pour l'appariement des données hydrographiques historiques avec des données récentes prend en compte la différence des précisions planimétriques qui peut être importante entre ces données de dates et de modes d'acquisition très différents. Cette approche commence par la détection et la hiérarchisation des « strokes » des cours d'eau (qui sont des structures naturelles de bonne continuité) dans les deux réseaux hydrographiques à apparier. Les niveaux hiérarchiques des strokes sont utilisés comme critère pour la sélection des candidats entre les deux jeux de données. La comparaison des strokes est réalisée ensuite selon des critères

géométriques et toponymiques. En outre, cette approche met à profit un critère de contexte spatial en calculant la distance la plus courte entre chaque stroke et les lieux nommés ponctuels figurant sur la même carte, préalablement appariés à ceux figurant sur la carte à l'origine du deuxième jeu de données. Dans cette approche, l'utilisation des différents critères d'appariement, y compris la prise en compte des distances avec les toponymes proches des cours d'eau, montre un grand potentiel pour réduire l'effet de l'imprécision géométrique et améliorer ainsi les résultats d'appariement. Cependant, cette approche reste restreinte aux données linéaires qui forment des réseaux. En outre, le critère de contexte rajouté dans cette approche nécessite des données supplémentaires fiables et déjà appariées et qui sont issues des mêmes sources cartographiques que les données à apparier, ce qui constitue des conditions difficiles à reproduire dans tous les cas.

Dans le même type d'approches qui prennent implicitement en compte les connaissances sur les hétérogénéités géométriques des données, (Bel Hadj Ali, 2001) a proposé une approche adaptée aux données surfaciques. Les objectifs principaux de cette approche sont de permettre la mise à jour ou le contrôle de la qualité d'un jeu de données en le comparant à un jeu de données de référence. Cette approche montre un potentiel important pour la prise compte des différences de modélisation géométrique et de granularité entre les jeux de données. Ceci est réalisé par un calcul d'intersection des deux jeux de données pour créer un ensemble de liens d'appariements probables. Les paires candidates sont ensuite filtrées : on supprime les paires parasites dont la distance surfacique est supérieure à un seuil fixé. Une matrice d'association est ensuite calculée à partir de ces paires pour détecter des liens de cardinalité 1:1 et ceux de cardinalité n:m. L'utilisation de cette matrice d'association permet de prendre en compte certains aspects de la différence de modélisation géométrique et de règles de saisie, notamment la différence de règles de sélection ou de décomposition. Par exemple l'approche permet de détecter des relations de cardinalité 1:n entre les objets qui représentent des bâtiments dans deux bases de données si les géométries sont agrégées par un critère d'adjacence dans la première base, alors qu'elles sont représentées distinctivement pour chaque bâtiment dans la deuxième base.

Algorithme d'appariement sans a priori sur les données géographiques en entrée

Contrairement aux approches de la partie précédente, certaines approches de l'état de l'art proposent une formalisation des connaissances sur les données pour pouvoir les prendre en compte dans le processus d'appariement. Il peut dans ce cas s'agir d'une explicitation des connaissances que l'expert possède ou peut acquérir sur les données et schémas qui décrivent les données.

Dans ce cadre, (Olteanu, 2008) propose un algorithme qui s'appuie sur l'utilisation de la théorie des fonctions de croyance (Dempster,1968 ; Shafer, 1976). Il définit pour chaque critère utilisé dans l'appariement trois fonctions, liées aux mesures utilisées pour chaque critère, qui précisent quand est-ce que ce critère est décisif positivement, quand est-ce qu'il est décisif négativement et quand est-ce qu'il est neutre pour l'appariement. Les critères sont ensuite combinés et les couples de candidats sont comparés entre eux. Cette approche est applicable pour n'importe quelle primitive géométrique et entre des jeux de données ayant des niveaux de détails proches ou différents. L'une des limites générales de cette approche est sa difficulté de paramétrage et son temps de calcul important rendant son passage à l'échelle très difficile. Cette même approche a été adaptée par (Duménieu, 2015) pour l'appariement de plusieurs sources de donnée, dans le cadre du suivi de l'évolution du réseau de rues parisiens au 19^{e} siècle à l'aide de de données géographiques issue de

plans anciens. Dans ce travail, le processus d'appariement est appliqué à plusieurs sources de données à la fois. Un modèle de représentation nommé « graphe géohistorique », basé sur les modèles classiques des données spatio-temporelles, a été proposé dans ce travail pour la représentation de données des différentes sources. Le but de l'appariement dans ce cas est d'arriver à combiner les données des différentes sources en un hypergraphe géohistorique. Les fonctions de croyance ne servent pas uniquement à décider de la création ou non de liens d'appariement, mais aussi à définir des relations de filiation entre les observations géohistoriques issues des différentes sources. Bien que l'utilisation de la théorie des fonctions de croyance permet d'expliciter, la connaissance de l'expert des données sur l'incertitude des différents critères d'appariement, notamment le critère géométrique, son utilisation reste limitée par rapport au problème d'hétérogénéité géométrique interne aux sources de données. En effet, la configuration de ces fonctions correspond à une connaissance générale sur les jeux de données impliqués et ne prend donc pas en compte la spécificité de chaque géométrie.

Dans un cadre d'appariement multicritère également, pour la propagation des mises à jour de bases de données géographiques, (Wang et al., 2015) propose une approche qui s'appuie sur l'apprentissage d'un modèle de réseau de neurones pour agréger les différentes valeurs de similarité calculées entre deux objets géographiques. Ce travail traite le cas des données surfaciques et exploite des critères d'appariement liés à leurs caractéristiques géométriques: la localisation, l'aire, la direction et la longueur. Selon les auteurs, l'utilisation du modèle de réseaux de neurones appris permet d'éviter la subjectivité de la fixation des poids des différents critères par l'expert des données. Le modèle est appris à partir d'un échantillon puis utilisé dans une approche d'appariement bidirectionnelle, *c.-à-d.* une approche constituée de la combinaison de deux phases d'appariements : une première phase directe entre le jeu de données à jour et le jeu de données à mettre à jour, et une deuxième dans le sens inverse. Pour gérer les mises à jour, les liens d'appariement sont interprétés selon leur cardinalité pour détecter les modifications (*ex.* un appariement de cardinalité 1 :0 est interprété comme une insertion de l'objet en question dans le jeu de données à mettre à jour). L'utilisation de l'apprentissage d'un modèle de réseau de neurones représente une manière d'acquérir et expliciter les connaissances sur l'hétérogénéité géométrique des données, ce qui constitue une piste intéressante pour nos problématiques. Cependant, l'approche reste plus adaptée aux jeux de données ayant chacun un contenu homogène, puisque le modèle de réseau de neurones appris est utilisé de la même manière partout entre les deux jeux de données.

Pour avoir un contenu cohérent lors d'un appariement de données géographiques, une cohérence entre les schémas des bases de données est nécessaire. En effet, l'alignement des schémas de bases de données géographiques est une étape aussi importante que l'appariement des objets géographiques dans un cadre d'intégration de données. Connaître les équivalences ou les spécialisations des classes et des attributs entre les différentes bases de données à apparier est primordial pour savoir quels objets on peut comparer et selon quels critères. Expliciter les schémas de données et les relations entre eux fournit une connaissance supplémentaire pour paramétrer le processus d'appariement. Plusieurs approches ont été proposées pour l'appariement des schémas des bases de données géographiques dans le but de les intégrer. (Ressler et al., 2009) propose une approche sémantique semi-automatique qui utilise des ontologies pour déterminer les classes similaires, et des règles métier pour automatiser la fusion de deux jeux de données géographiques. L'algorithme de cette approche se compose de trois étapes principales: la première consiste à extraire, à partir des deux sources, les concepts auxquels les entités géographiques peuvent

appartenir en proposant des équivalences entre les valeurs des attributs et les concepts d'un vocabulaire général généré a priori par les auteurs. La deuxième étape consiste à présenter une interface pour ces équivalences afin de les faire valider (ou décliner) par un expert du domaine, ce qui permet d'associer les objets aux concepts définis. La troisième étape consiste à inférer des règles métiers à partir des liens entre les objets et les concepts du vocabulaire utilisé. Ces règles associent les classes équivalentes entre les deux jeux de données. L'utilisation d'un vocabulaire général pour typer les données dans cette approche permet d'automatiser la configuration de la tâche de fusion. Cependant, cette configuration ne prend pas en compte les différentes hétérogénéités liées aux géométries des données.

(vidal et al., 2009) propose une approche qui s'appuie plutôt sur l'alignement des schémas qui structurent les données géographiques comme moyen d'intégration virtuelle de bases de données géographiques. Il ne s'agit donc pas de chercher des liens explicites d'appariement entre les objets de bases de données, mais d'un système qui permet un accès simultané à ces bases. Dans cette approche, des ontologies d'application sont générées à partir des schémas des deux bases de données. Tout d'abord ces ontologies d'application sont alignées à une ontologie du domaine. Ces alignements sont utilisés pour la réécriture des requêtes exprimées selon les termes de l'ontologie du domaine en sous-requêtes exprimées selon les termes des ontologies d'application. Les alignements sont également utilisés pour gérer la fusion des résultats des sous-requêtes en éliminant les redondances et en les réécrivant les résultats selon les termes de l'ontologie du domaine. L'utilisation d'une ontologie du domaine permet donc de lier virtuellement les objets équivalents entre les deux sources. Cette approche permet donc de remédier aux problèmes d'hétérogénéités des schémas, mais ne fournit pas un de solution pour gérer les hétérogénéités géométriques.

L'utilisation d'ontologies de domaine pour l'intégration des données s'est montrée efficace dans l'approche proposée par (Uitermark et al., 2001). Cette approche à deux phases utilise les relations déduites entre les classes des ontologies de deux jeux de données afin de filtrer les résultats d'appariements entre les instances de ces jeux. En première phase, une ontologie du domaine ainsi que les connaissances sur les spécifications des deux jeux de données sont utilisées afin de créer « un modèle de référence ». Ce modèle définit les relations entre les classes des ontologies des deux bases de données. La deuxième phase consiste à chercher les correspondances entre les instances des deux bases par un appariement géométrique, puis filtrer ces correspondances en s'appuyant sur les relations entre les classes créées dans la première phase. Dans cette approche, les connaissances sur les spécifications des données concernées sont prises en compte pour l'explicitation des relations entre les schémas qui sont utilisées dans le filtrage à postériori des résultats d'appariement. Elles ne sont donc pas utilisées pour mieux paramétrer la comparaison géométrique. De plus, les connaissances sur les spécifications des données sont seulement analysées pour définir les relations entre les classes. Elles ne sont donc pas vraiment formalisées, et ne peuvent pas être prises en compte d'une manière automatique pour une autre application que celle présentée ici.

(Gesbert, 2005) et (Abadie, 2012 se sont focalisés sur la formalisation des spécifications de bases de données géographiques afin de permettre l'intégration de celles-ci. (Gesbert, 2005) propose un langage formel qui permet d'exprimer les relations entre les concepts d'une ontologie du domaine et les classes des schémas des bases de données à intégrer en s'appuyant sur leurs spécifications. (Abadie, 2012) propose l'utilisation du langage standard du Web sémantique OWL2 pour formaliser les spécifications des bases de données géographiques. Cette formalisation est exploitée dans des

approches d'appariement de schémas dans un but d'intégration virtuelle des bases de données géographiques. En plus d'être utile pour l'alignement des schémas, formaliser les spécifications des bases de données géographiques présente un intérêt particulier pour la prise en compte des hétérogénéités dans la comparaison de leurs géométries. En effet, formaliser des spécifications de données telles que la modélisation géométrique ou les critères de sélection et de décomposition peut être un moyen d'automatiser certains choix de paramétrage pour la comparaison des géométries. La formalisation des spécifications est cependant instanciée pour chaque classe de données dans une base de données géographiques. Ceci suppose une homogénéité des spécifications pour chaque classe à l'intérieur d'une même source de données, ce qui constitue une limite face à l'hétérogénéité géométrique interne des jeux de données du Web.

2.2 Interconnexion des Linked Data : concept, approches et outils

Créer des liens constitue un enjeu crucial pour atteindre la vision d'un Web de données. Il est à l'origine de l'émergence de nombreux outils et approches d'interconnexion. Le développement de ces outils et approches trouve une bonne base dans l'héritage conséquent d'approches issues de domaines qui, historiquement, s'intéressent à des problématiques similaires ou identiques à celles rencontrées avec l'interconnexion de données liées, tels que le couplage d'enregistrement en fichiers et en base de données (record linkage *cf*. (Winkler, 2006 ; Christen, 2012)), la résolution d'entités nommées et la détection de doublons (*entity resolution*, *duplicate detection cf*. (Christen, 2012)) ainsi que les travaux sur l'alignement d'ontologies (*ontology matching cf*. (Euzenat et Shvaiko, 2007)). Des travaux tels que (Scharffe et Euzenat, 2010; Scharffe et al., 2011; Ferrara et al., 2013 ; Nentwig et al., 2015 ; Achichi et al., 2016)) proposent des états de l'art des différents outils et approches d'interconnexion de données sur le Web. Nous présentons dans cette section les différentes étapes qui constituent un processus d'interconnexion. Nous exposons ensuite les outils d'interconnexion susceptibles de remplir, même partiellement, nos objectifs.

2.2.1 Étapes d'un processus d'interconnexion

L'interconnexion peut être formalisée simplement comme un processus qui prend en entrée deux jeux de données (souvent une source et une cible), pour donner en sortie un ensemble de liens de correspondance entre les ressources de ces jeux. Plus en détail, le flux de travail d'un processus d'interconnexion est constitué de trois étapes élémentaires que distingue (Ferrara et al., 2013 ; Nentwig et al., 2015 ; Achichi et al., 2016): le prétraitement, la mise en correspondance (*matching*) , le post-traitement. La Figure 2.4 représente le schéma général de ce processus d'interconnexion.

Ces trois étapes sont précédées par une configuration (fournie par l'utilisateur) qui consiste à choisir les paramètres nécessaires à l'interconnexion comme les sources de données concernées, les propriétés à comparer, les mesures de distance à utiliser pour la comparaison, les seuils de tolérance pour les distances calculées, le type des liens à créer…etc. Ce paramétrage peut être codé en dur au cœur de l'outil d'interconnexion, comme il peut être exprimé sous la forme d'un ensemble de spécifications codées dans un langage compréhensible par l'outil d'interconnexion, et donc fourni en entrée pour chaque exécution. Dans d'autres cas, certains de ces paramètres ne sont pas fournis en amont, mais peuvent être extraits d'une manière plus automatique (*ex.* en exploitant un fichier d'alignement des schémas qui décrivent les données).

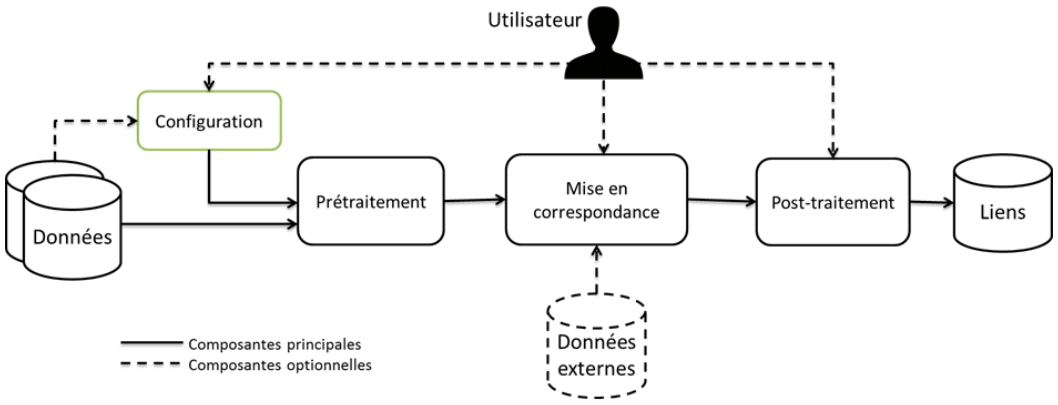


Figure 2.5 Schéma général d'un processus d'interconnexion adapté du schéma proposé par (Ferrara et al., 2013).

Phase de prétraitement

La phase de prétraitement est principalement une étape de préparation et de chargement de données. L'un des objectifs principaux pour cette étape est d'optimiser le déroulement de l'étape suivante de mise en correspondance. Le prétraitement participe donc à réduire l'espace de recherche pour minimiser le nombre de comparaisons dans la phase suivante. Étant donné que l'interconnexion ne s'applique pas à toute ressource se trouvant dans les jeux de données en entrée, une restriction est généralement appliquée pour sélectionner seulement des ressources qui respectent certains critères (*ex.* sélectionner seulement les instances d'une classe précise). Une restriction sur les propriétés utiles peut être également appliquée pour ne charger que celles qui sont susceptibles d'être nécessaires dans la phase de mise en correspondance. Pour automatiser cette étape, le prétraitement comprend, dans certaines applications, des approches de détection de clés ou de quasi-clés. Comme expliqué précédemment (*cf.* section 1.4.1), des approches telles que celles proposées par (Pernelle et al., 2013; Atencia et al., 2012; Symeonidou et al., 2014) permettent de détecter les propriétés qui identifient les ressources d'un jeu de données d'une manière unique (*c.-à-d.* les clés), ou celles dont la combinaison avec d'autres propriétés permet de le faire (*c.-à-d.* les quasi-clés). La restriction des propriétés participe à l'optimisation de l'espace utilisé pour charger les données. Pour optimiser le temps de traitement en réduisant encore le nombre de comparaisons nécessaires entre ressources dans la phase suivante, le prétraitement peut inclure également des techniques de partitionnement et de filtrage.

Des opérations de transformation des valeurs des propriétés sont susceptibles d'être utilisées pour charger les données sous des formes comparables. Par exemple, pour une éventuelle comparaison de valeurs représentant des dates, ces valeurs doivent être mises dans un même format pour éviter toute confusion liée à l'ordre des jours et des mois par exemple. Les opérateurs de transformation de chaînes de caractères sont très utilisés dans ce cadre. Il peut s'agir par exemple d'un simple changement de casse, d'une concaténation de plusieurs valeurs (*ex.* concaténer les valeurs d'attributs de numéro, de nom de voie et de code postal pour reconstituer une adresse), d'un remplacement ou d'une suppression d'une partie de la chaine de caractères, etc. Effectuer la transformation sur les valeurs des propriétés lors du chargement des données permet d'alléger le traitement des données dans la phase suivante.

Phase de mise en correspondance

La phase de mise en correspondance, dite de « *matching* », représente le cœur du processus d'interconnexion. Elle consiste, dans le cas le plus simple, à comparer les descriptions des ressources afin d'en tirer la décision de créer ou non des liens entre elles. La technique habituelle compare les valeurs des propriétés en utilisant des mesures pour calculer leur similarité. Une technique d'agrégation de ces valeurs de similarité peut être nécessaire dans ce cas. Il s'agit donc d'un processus multicritère de mise en correspondance. Les mesures de distance ou de similarité utilisées dans ce cas sont semblables, voire identiques, à celles utilisées pour l'appariement de données géographiques ou n'importe quel autre domaine de mise en correspondance de données. Les mesures de distance géométriques présentées dans la (section 2.1.1) sont également utilisables dans ce cas, même si elles restent implémentées pour quelques approches ou outils d'interconnexion seulement. Pour les autres types de propriétés, on dispose d'un état de l'art conséquent de mesures de distance ou de similarité.

Comparer des propriétés non géométriques, quantitatives (*ex.* population, date) représente l'un des cas les plus simples. Comparer des propriétés de ce genre revient à calculer une différence numérique ou un ratio entre les valeurs, ou à chercher une simple égalité. Les mesures de distance ou de similarité utilisées doivent néanmoins prendre en compte les unités de mesure des valeurs (*ex.* unité de longueur) et le format des valeurs (*ex.* format de date).

Les propriétés qualitatives qui représentent des valeurs non quantifiables stockent l'information sous forme de chaînes de caractères (*ex.* nom, description, nature). De nombreuses mesures de distance ou de similarité entre valeurs de chaînes de caractères qui ont été proposées dans la littérature sont le fruit de travaux de plusieurs communautés. (William W. Cohen et al, 2003) propose une comparaison des différentes mesures proposées dans la littérature pour la comparaison des chaînes de caractères. (Elmagarmid, 2007) décrit également les différentes mesures de similarité ou de distance utilisées pour la détection des enregistrements en doublons dans les bases de données. On identifie principalement cinq types de mesures dans ce cas : les mesures basées sur les caractères, les mesures basées sur les mots, les mesures basées sur la phonétique, les mesures basées sur la sémantique et les mesures hybrides.

Les mesures basées sur les caractères considèrent la valeur d'une propriété comme une seule chaîne de caractères à comparer sans prendre en compte sa décomposition lexicale. Ce type de mesure est mieux adapté à la comparaison de valeurs de propriétés qui sont susceptibles de contenir un seul mot ou des expressions simples où l'ordre des mots est important (*ex.* nom, nature). Dans cette catégorie, les distances dites « d'édition » reposent sur le nombre d'altérations minimum pour que deux chaînes de caractères soient égales. La distance de *Levenshtein* (Levenshtein,1966) est l'une des distances d'édition les plus utilisées. Elle calcule le nombre minimal d'ajouts, de suppressions et de remplacements de caractères pour passer d'une chaîne à une autre. Dans ce cas la distance entre les mots « route » et « routes » est 1, et entre les mots « sentier » et « chantier » est 3. La distance de *Levenshtein* peut être normalisée en divisant le coût d'édition par la plus grande des longueurs des deux chaînes. D'autres distances d'édition existent comme la distance de *Smith-Waterman* (Smith et Waterman, 1981) et sa variante normalisée *Monger-Elkan* (Monge et Elkan, 1996). Il existe d'autres mesures basées sur les caractères, mais qui ne sont pas considérées comme des distances d'édition telles que la mesure de similarité *Jaro* (Jaro, 1976) et sa variante *Jaro-Winkler* (Winkeler, 1991). La

mesure de *Jaro* est normalisée. Elle est basée sur le nombre de caractères communs ayant les mêmes positions dans les deux chaînes ainsi que le nombre de transpositions entre ces chaînes. La variante *Jaro-Winkler* reprend le même principe en favorisant les chaînes de caractères qui partagent le même préfixe. D'autres mesures qui ont été proposées dans la littérature peuvent servir pour comparer des chaînes en se basant sur les caractères comme la distance de *Hamming* (Hamming, 1950) ou la distance (et sa similarité associée) *N-grams* (Kondrak, 2005).

Alternativement, deux chaînes de caractères peuvent être vues comme deux ensembles d'unités lexicales (sacs de mots). Via un processus dit de « *tokenisation* », les chaînes de caractères peuvent être découpées en ensembles qu'on peut comparer grâce aux mesures basées sur les mots. Ces mesures semblent plus efficaces pour comparer les chaînes de caractères portant de longues phrases (*ex.* description d'un objet) ou des textes. Dans cette catégorie, la mesure de similarité de *Jaccard* (Jaccard, 1901) représente un cas classique. C'est une mesure normalisée basée sur le rapport entre le nombre de mots communs entre deux ensembles de mots et le nombre de mots de l'union de ces deux ensembles. Entre deux ensembles de mots S et T la similarité de *Jaccard* est calculée comme suit :

$$Sim_{jacc} = \frac{|S \cap T|}{|S \cup T|} \qquad 2.6$$

Lorsqu'elle est appliquée pour comparer des chaînes de caractères, cette mesure ne prend donc pas en compte l'ordre des mots dans les chaînes. Similairement à la mesure de Jaccard, la mesure de similarité de *Sørensen-Dice* est basée sur le nombre de mots communs entre deux ensembles de mots S et T, et se calcule comme suit :

$$Sim_{sor} = \frac{2|S \cap T|}{|S| + |T|} \qquad 2.7$$

Une autre mesure entre ensembles de mots est la mesure de similarité *Cosine* qui est basée sur la fréquence de répétition des mots communs entre deux ensembles dans chacun des ensembles.

Les mesures basées sur la phonétique s'appuient sur la comparaison des prononciations des mots pour décider de leur ressemblance. *Soundex* (Russell, 1918) est l'exemple le plus connu des algorithmes phonétiques. Il permet d'encoder les mots qui se ressemblent au niveau de la prononciation dans des chaînes de caractères (codes) similaires. L'idée principale est de supprimer certaines lettres comme les voyelles, puis pour le reste des lettres, encoder avec un même chiffre les lettres qui partagent le même point d'articulation. Sa version originale a été conçue pour la langue anglaise, mais peut être adaptée aux autres langues notamment le français. *Metaphone* est un algorithme d'encodage phonétique qui a été proposé par (Philips, 1990) comme une amélioration de *Soundex. Double-Metaphone* est une version encore plus avancée proposée par (Philips, 2000) et qui comprend des améliorations comme l'encodage multiple pour les noms qui ont plusieurs prononciations possibles et la possibilité d'adapter l'algorithme aux langues autres que l'anglais. D'autres algorithmes d'encodage phonétique comme *NYSIIS* (Taft, 1970) et ONCA (Gill, 1997) ont été proposés pour surmonter les limites de *Soundex.*

Au-delà de sa structure syntaxique, une chaîne de caractères porte un sens (une sémantique). Il existe une autre famille de mesures de distance ou de similarité qui s'appuient sur la sémantique des

mots. L'une des plus importantes complexités lors de la comparaison de deux valeurs de propriétés est de savoir si elles portent le même sens même si elles sont syntaxiquement différentes. Ceci survient par exemple quand on utilise des synonymes pour exprimer les deux valeurs, si une valeur représente un concept plus général que l'autre, ou si les deux valeurs sont exprimées dans deux langues différentes. Pour cela on utilise ce qu'on appelle une mesure de distance sémantique entre concepts appartenant à une taxonomie ou un dictionnaire. L'avancée des travaux du domaine de la représentation de connaissances, notamment ceux liés au Web sémantique, a engendré la création d'une multitude d'ontologies générales ou liées à des domaines spécifiques, ainsi que le développement de mesures de distance entre concepts qui s'appuient sur ces ontologies. Dans ce cadre, *Wordnet* (Miller, 1995 ; Fellbaum, 1998) est l'une des taxonomies les plus utilisées. Son but est de regrouper et classer, selon la sémantique, les mots (verbes, noms, adjectif et adverbes) de la langue anglaise. Il existe cependant des versions de cette taxonomie dans d'autres langues. Des travaux comme (Resnik, 1995) et (Hirst et St Onge, 1998), par exemple, se basent principalement sur l'utilisation de *Wordnet* dans la proposition de mesures de distance sémantiques. (Wu et Palmer, 1994} propose également une distance basée sur la profondeur des concepts et leur plus proche ancêtre commun dans une hiérarchie de concepts afin de quantifier leur distance sémantique.

Les familles de mesures de distance ou de similarité peuvent être utilisées séparément les unes des autres. Or en pratique, pour combiner les avantages de ces mesures, des mesures hybrides qui mélangent plusieurs techniques des catégories précédentes sont utilisées pour comparer des chaînes de caractères. On peut par exemple utiliser une mesure qui marie une technique basée sur les mots (ex *Jaccard*) avec une technique basée sur les caractères (*ex. Levenstein*). Dans ce cas si on dispose de deux chaînes de caractères et qu'on réussit à les découper (tokeniser) en deux ensembles de mots, l'intersection entre ces deux ensembles ne se calcule pas en cherchant les mots exactement égaux (comme dans le cas basique de *Jaccard*), mais cherchant ceux qui ont une distance de caractères (*ex. Levenstein*) inférieure à un certain seuil. (Monge et Elkan, 1996) propose une mesure de similarité dans ce genre, en combinant l'utilisation de la *tokenisation* avec une mesure de similarité basée sur les caractères.

Le déroulement de cette phase de mise en correspondance peut changer drastiquement d'une approche à l'autre (*cf.* section 2.2.2) : d'autres techniques peuvent être introduites dans cette phase, telles que la comparaison des structures topologiques des ressources, la propagation des liens, l'apprentissage pour la validation des liens, etc. Dans certains cas, cette phase nécessite une interaction avec l'utilisateur afin de garantir son bon déroulement. Selon les techniques utilisées dans cette phase, chacun des liens obtenus peut être associé à une valeur qui représente son score de confiance. Toutes les techniques utilisées dans cette phase sont définies par les paramètres qui sont choisis dans la configuration initiale de l'interconnexion ou extraites d'une manière automatique dans la phase de prétraitement.

Phase de post-traitement

Le but de cette étape consiste principalement à filtrer les liens résultant de l'étape précédente en respectant des critères comme : pour chaque ressource de la source ne garder que les liens ayant un score de confiance supérieur à un seuil d'acceptation, ne garder que le lien ayant le plus grand score si la cardinalité des liens est de 1:1 ou garder les n premiers liens selon le score de confiance si la cardinalité des liens est de 1:n. L'un des objectifs principaux de cette phase est la validation de liens

résultants. Cette validation peut se faire manuellement ou semi automatiquement (*ex.* par apprentissage de règles de validation).

2.2.2 Approches et outils d'interconnexion de Linked Data

Les approches et techniques de l'état de l'art sur l'interconnexion de ressources diffèrent les unes des autres sur plusieurs plans. (Scharffe et al., 2011; Ferrara et al., 2013), par exemple, classifie les techniques utilisées selon trois critères principaux : le niveau de granularité sur lequel elles opèrent (au niveau des jeux de données, des ressources ou des valeurs de propriétés), le type d'information utilisée (les données ou les connaissances qui décrivent les données), ainsi que l'origine des informations utilisées (interne ou externe). (Nentwig et al., 2015) propose une comparaison fonctionnelle de quelques outils d'interconnexion de l'état l'art. Cette comparaison porte sur des aspects comme : la source de données (RDF ou autres), le type de configuration initiale (manuelle ou automatique par apprentissage, type d'agrégation, etc.), les mesures de similarité utilisées, l'utilisation de la comparaison sémantique, le type de post-traitement s'il existe, les aspects d'optimisation du temps d'exécution (indexation, filtrage, exécution parallèle), etc.

Les différences entre les approches et outils existants résident dans l'angle sous lequel ils abordent le problème d'interconnexion et donc les différentes techniques qui y sont mises en œuvre. Nous mettons l'accent ici sur certains outils et approches qui sont susceptibles de répondre aux objectifs avancés dans la section 1.6 d'avoir un outil générique capable d'adapter dynamiquement la configuration du processus d'interconnexion pour comparer les géométries en prenant en compte les hétérogénéités qui peuvent exister entre elles. Nous présentons d'abord les approches qui implémentent des processus génériques indépendants d'un domaine d'application spécifique. Nous verrons ensuite les différentes approches à base de connaissances proposées pour améliorer le processus d'interconnexion. Nous nous intéresserons enfin aux approches qui proposent une adaptation du processus d'interconnexion aux caractéristiques des données impliquées.

Des approches et outils génériques pour l'interconnexion des données

L'accroissement constant du nombre de ressources publiées sur le Web de données a sucité la proposition de nombreux outils et approches qui visent à résoudre le problème de l'interconnexion des données. Certains d'entre eux ont ciblé des données de domaines d'application spécifiques tels que l'outil **LD-Mapper**(Raimond et al., 2008} implémenté pour fonctionner avec des données du domaine de la musique, ou l'outil **RKB-CRS**(Jaffri et al.,2011} pour trouver des équivalences liées au domaine des publications scientifiques et universitaires. Ces outils restent restreints à leurs domaines d'application, car ils utilisent des techniques qui y sont spécifiques comme l'analyse de co-auteurs dans le cas **RKB-CRS**.

D'autres outils ont visé à implémenter des processus beaucoup plus génériques capables de fonctionner avec des données issues de domaines quelconques et dans des contextes très variés. Nous nous intéressons particulièrement à ce cas, car l'un de nos objectifs et de pouvoir proposer une approche qui bénéficie le plus possible des avantages de ces outils génériques tels que la possibilité d'être utilisés pour le plus de cas d'application possibles et les techniques d'optimisation du temps et de l'espace de travail. Les outils génériques les plus connus par la communauté d'interconnexion de

données sont **LIMES**[65](Ngonga Ngomo et Auer, 2011) et **Silk**[66] (Silk Link Discovery Framework) (Volz et al., 2009).

LIMES est un outil d'interconnexion générique semi-automatique, car il nécessite en entrée un script qui rassemble les spécifications d'exécution le plus souvent créé manuellement. La technique principale utilisée pour la mise en correspondance des ressources est la comparaison multicritère. Plusieurs types de mesures de distance entre les ressources sont inclus dans l'outil, comme les mesures entre chaînes de caractères, les mesures de distance entre géométries, les mesures de distance temporelle, etc. Une fonction d'agrégation (*ex.* maximum) est utilisée pour combiner les scores calculés par les mesures de distance choisies.

De son côté, **Silk** est également un outil générique, multicritère et semi-automatique d'interconnexion. Il est conçu principalement pour traiter des données liées à partir de fichiers RDF ou de points d'accès Sparql, mais peut traiter également d'autres formats tels que les fichiers CSV ou les bases de données SQL. Des fonctions de transformation des valeurs de propriétés (remplacement de caractères, combinaison, normalisation, *etc.*) sont proposées dans cet outil pour permettre la transformation des valeurs de propriétés sous des formes comparables. Silk utilise également la technique de comparaison multicritère : un ensemble de mesures de distance sont mises à disposition pour comparer les chaînes de caractères, les valeurs numériques, les géométries, les dates, etc. Différentes fonctions d'agrégation sont disponibles également telles que la moyenne pondérée, le maximum, le minimum, la moyenne géométrique, etc.

En plus d'implémenter le processus d'interconnexion avec ses étapes présentées dans 2.2.1, **LIMES** et **Silk** intègrent donc des mesures pour la comparaison géométrique. En effet, dans le cadre de **Silk**, (Smeros et Koubarakis, 2016) proposent un ensemble de mesures de distance spatiales et de vérification de relations topologiques. Dans le cadre de **LIMES**, pour permettre l'utilisation des distances entre géométries d'une manière efficace, l'approche d'optimisation ORDCHID (Ngonga Ngomo, 2013) a été proposée pour minimiser le temps de comparaison géométrique dans le processus d'interconnexion. Dans le cadre de **LIMES** également, une approche RADON (Ahmed Sherif et al., 2017) a été proposée pour permettre à l'outil de découvrir des relations topologiques entre les ressources. L'intérêt porté par ce genre d'outils à l'utilisation de la géométrie pour l'interconnexion augmente leur potentiel pour répondre à nos objectifs.

Il existe cependant d'autres propositions qui se sont appuyées sur l'utilisation de la comparaison des géométries comme critère principal d'interconnexion bien qu'elles ne soient pas vraiment génériques. Nous faisons référence ici à des approches comme celle proposée par (Hahmann et Burghardt, 2010) restreinte seulement à l'interconnexion de données géoréférencées par des géométries ponctuelles, ou l'approche proposée par (Salas et Harth, 2011) restreinte seulement à l'interconnexion de données géoréférencées par des géométries surfaciques. L'outil **Zhishi.links**(Niu et al., 2011} utilise également la comparaison de coordonnées géographiques en plus des labels comme un critère d'interconnexion. (Vilches-Blázquez et al., 2012) propose une approche composée d'un ensemble d'heuristiques pour l'interconnexion de données géographiques converties en RDF. Les heuristiques développées dans ce travail permettent la comparaison : des noms locaux des URIs

[65] http://aksw.org/Projects/LIMES.html
[66] http://silkframework.org/

des ressources entre les deux jeux de données, des labels et leurs langues entre les deux jeux de données, des labels des ressources dans un jeu avec les noms locaux des URIs dans l'autre, et des géométries des ressources entre les deux jeux de données. Les comparaisons sont effectuées grâce à une variété de mesures de similarité et d'égalité. Les résultats des heuristiques sont ensuite combinés pour retourner à l'utilisateur deux fichiers: le premier contient les liens owl:sameAs entre les entités considérées comme équivalentes alors que le deuxième contient une liste de paires de ressources candidates à valider par un expert du domaine.

La réduction de la complexité et le passage à l'échelle constituent des caractéristiques importantes pour accroitre la généricité d'un outil d'interconnexion. Dans ce cadre, l'avantage principal de l'outil **Zhishi.links** est sa capacité à indexer les données et l'utilisation de techniques flexibles qui passent à l'échelle grâce à une architecture distribuée. L'outil **OKKAM**(Bouquet et al., 2008) gère également la réduction de la complexité temporelle en utilisant également une architecture distribuée de serveurs (dits « de nom d'entités ») qui contiennent chacun de ressources équivalentes : l'ajout d'une nouvelle ressource se réalise en comparant sa description, d'une manière parallèle, avec les descriptions des ressources contenues dans les serveurs. L'un des avantages principaux de **LIMES** réside dans les techniques d'optimisation qui y sont employées qui réduisent la complexité temporelle du processus d'interconnexion en réduisant son nombre de comparaisons. L'approche ORCHID(Ngomo, 2013) proposée dans le cadre de **LIMES,** et qui vise à réduire le nombre de comparaisons entre géométries lors de l'utilisation de la distance de *Haussdorff* et de la distance *orthodromique*, en est un exemple concret. **Silk** implémente également dans sa phase de prétraitement des techniques pour réduire l'espace de recherche de liens. Pour éviter une comparaison globale qui reviendrait à effectuer un produit cartésien entre le jeu de données source et le jeu de données destination, **Silk** implémente une méthode d'indexation pour partitionner les ressources dans des blocs selon la similarité des valeurs de propriétés concernées par la comparaison. Avec chaque mesure de distance, une fonction d'indexation est donc implémentée pour permettre d'assigner à chaque ressource le numéro de bloc approprié à la valeur de sa propriété concernée par la comparaison. Dans le cas d'une comparaison multicritère, Silk introduit la technique de « *blocking* » multiple qui agrège les différentes indexations (Isele et al., 2011) : selon la fonction d'agrégation utilisée, une ressource peut être assignée à plusieurs blocs. Cette technique permet de réduire la complexité temporelle de la phase de mise en correspondance d'une complexité quadratique à une complexité linéaire tout en gardant le même rappel. **Silk** et **LIMES** sont proposés également sous la forme d'une implémentation en architecture distribuée (MapReduce) pour le passage à l'échelle dans le cas de jeux de données très volumineux.

Les outils génériques se distinguent également par leur mode de configuration. À l'inverse des outils adaptés à des domaines spécifiques, dont le code nécessite d'être changé pour chaque cas d'application, les outils génériques semi-automatiques jouissent d'une flexibilité de configuration permettent de les adapter aux données à interconnecter. C'est le cas de **RDF-AI** qui est un prototype d'outil proposé par (Scharffe et al., 2009) à des fins d'interconnexion ou de fusion de données. Les paramètres d'exécution de cet outil sont fournis en entrée sous la forme de fichier XML qui précise où trouver la structure ontologique des données, quelles sont les transformations à effectuer sur les valeurs des propriétés et quelle technique de mise en correspondance utiliser pour chaque type de ressources. Des fichiers, séparés des données à interconnecter, qui spécifient les structures ontologiques de chaque source doivent être fournis. Ce sont ces structures ontologiques qui précisent quelles sont les ressources à interconnecter et quelles sont les propriétés à comparer. Un

autre fichier de configuration peut être fourni en outre pour préciser des paramètres de post-traitement tel que le seuil de score de création des liens. **Silk** nécessite également en entrée un script qui décrit, dans un langage spécifique (Silk-LSL), les spécifications du processus d'interconnexion. Ceci inclut les jeux de données en entrée, les critères de sélection des ressources à interconnecter dans chaque jeu de données, les critères de comparaison, le type des liens à créer, etc. C'est également le cas de l'outil **LIMES**. Ces outils sont dits semi-automatiques, car la configuration des règles de spécifications d'interconnexion est fournie par l'utilisateur de l'outil qui connaît ses données et leur structure. Nous verrons dans la suite comment les connaissances sur les données sont utilisées pour établir cette configuration d'interconnexion, permettant ainsi d'automatiser encore plus le processus d'interconnexion.

Utilisation de connaissances sur les données pour la configuration de l'interconnexion

Comme décrit précédemment, configurer le processus d'interconnexion revient à fixer des choix sur l'accès aux sources des données, la sélection des entités à interconnecter, les critères de comparaison et les mesures à utiliser, les méthodes d'agrégation, la cardinalité des liens, etc. Pour les outils d'interconnexion génériques, cette configuration peut être préparée indépendamment des outils grâce à des fichiers scripts. Établir cette configuration manuellement dans ce cas peut être considéré comme une explicitation des connaissances que l'utilisateur a sur ses données afin de paramétrer le processus d'interconnexion. Toutefois, certaines approches proposent de se passer en grande partie des connaissances de l'utilisateur en s'appuyant uniquement sur les connaissances formelles qu'on peut trouver au niveau ontologique ou celles extraites par apprentissage sur les données et les vocabulaires qui les décrivent. Nous nous intéressons particulièrement à ces approches, car les techniques appliquées dans ce cadre peuvent s'avérer utiles dans la prise en compte, dans un processus d'interconnexion, des connaissances sur les causes hétérogénéités géométriques.

Dans certains cas, ces connaissances formelles ne sont autres que les ontologies qui décrivent les données. **LN2R**(Saïs et al., 200) est un exemple d'approche complètement automatique grâce à l'utilisation d'une ontologie partagée qui décrit les données pour piloter le processus d'interconnexion. L'ontologie est en effet utilisée pour trouver les classes des ressources à interconnecter ainsi que leurs propriétés comparables. Grâce à la sémantique de cette ontologie, les dépendances entre les différentes ressources sont prises en compte pour calculer les valeurs de similarité finales pour chaque ressource.

Dans d'autres cas, les approches d'interconnexion s'appuient sur l'alignement des vocabulaires qui décrivent les sources de données utilisées comme des connaissances à partir desquelles on peut automatiquement tirer des règles de sélection de ressources et de comparaison de propriétés. C'est le cas par exemple de l'outil **Knofuss** qui permet de fusionner des sources de données structurées selon des vocabulaires différents. L'alignement produit entre les deux vocabulaires (par un outil d'alignement automatique de schémas externe à **Knofuss**) est traduit en requêtes Sparql qui permettent de sélectionner les ressources pertinentes et leurs propriétés et de les transformer d'un schéma à l'autre. Ainsi, les ressources sont comparées comme si elles étaient structurées dans un même vocabulaire. L'outil d'interconnexion **SERIMI**(Araujo et al., 2011) s'appuie sur une technique similaire, car il emploie un calcul du taux des propriétés partagées par les classes de ressources. Ça ressemble dans ce cas à un alignement des vocabulaires qui structurent les données, même si la

particularité de cet outil est que cette opération n'est pas réalisée en amont, mais pendant la phase de mise en correspondance.

D'autres outils appliquent des méthodes d'apprentissage pour extraire les règles de sélection et de comparaison de ressources. La construction des spécifications d'interconnexion d'une manière automatique par apprentissage a été proposée dans le cadre de l'outil **Knofuss** (Nikolov et al., 2008): en cas d'absence d'alignement de schémas entre les sources de données, un algorithme génétique est appliqué pour sélectionner les paires de propriétés discriminantes qui contiennent des valeurs comparables. Cette technique est utilisée également par l'outil **ObjectCoref**(Hu et al., 2011) qui, à partir d'un ensemble de liens d'apprentissage, cherche les paires « propriété – valeur » similaires entre les descriptions des ressources qui sont liées par ces liens. Dans une optique similaire, l'approche proposée par (Fan et al., 2014) découvre les correspondances entre les classes et entre les propriétés des différents jeux de données grâce à une méthode d'apprentissage non supervisée. Les alignements entre les propriétés permettent la construction de « patrons » d'interconnexion qui précisent les règles de spécifications d'interconnexion. Dans le cadre de **Silk**, (Isele et Bizer, 2011; 2012) proposent, alternativement à la configuration manuelle, l'apprentissage des règles d'interconnexion en utilisant des méthodes de programmation génétique. Similairement, dans le cas de **LIMES**, les méthodes EAGLE(Ngomo et Lyko, 2012), EUCLID(Ngomo et Lyko, 2013) et COALA(Ngomo et al., 2013) ont été proposées pour utiliser des algorithmes d'apprentissage (actif ou supervisé) sur les sources de données afin de générer des spécifications d'interconnexion entre elles. Dans le cadre de **LIMES** également, ROCKER est un opérateur proposé par (Soru et al., 2015) qui sert à détecter les clés qui peuvent servir pour l'interconnexion.

Dans le cadre des travaux d'alignement d'ontologies, où le processus de mise en correspondance est similaire à celui de l'interconnexion des données, (Duchateau et al, 2007) propose une approche qui remplace les fonctions d'agrégation habituelles par un arbre de décision. Ce dernier permet de planifier d'une manière flexible le choix des métriques de comparaison et leur combinaison. Les mêmes auteurs proposent dans (Duchateau et al, 2008) une adaptation de cette méthode aux besoins de l'utilisateur en utilisant une méthode d'apprentissage pour apprendre les combinaisons de mesures de comparaison les plus appropriées pour l'arbre de décision. L'utilisateur peut ainsi choisir l'arbre dont le compromis entre la qualité des résultats et les performances d'exécution répond le mieux à ses attentes. Dans le cadre des travaux d'alignement d'ontologies également, (Huza et al.,2006) propose un ensemble de critères de classification des méthodes d'alignement afin de permettre de sélectionner celles qui sont les mieux adaptées aux contextes des ontologies en entrée. Le système proposé dans ce travail se base sur un ensemble de règles de décision qui établissent des liaisons entre les critères de classification des contextes d'ontologies et les critères de sélection des méthodes d'alignement. Dans une approche similaire, (Mochol et Jentzch, 2008) proposent un système à base de règles pour la sélection des approches d'alignement les plus adaptées aux ontologies en entrée. L'originalité de cette approche réside dans l'utilisation de métadonnées qui décrivent les ontologies et de métadonnées qui décrivent les méthodes d'alignement comme entrée pour le système à base de règles. Plusieurs défis ressentent à relever pour le paramétrage automatique des outils d'alignement d'ontologies, notamment d'améliorer la combinaison des différentes approches d'alignement sélectionnées automatiquement (Shvaiko et Euzenat, 2013). Toutefois, les solutions proposées par ces outils constituent des pistes intéressantes sur l'utilisation des règles de décision basées sur les connaissances qui décrivent données afin d'adapter la tâche de mise en correspondance aux données à traiter.

Les techniques basées sur les connaissances qui décrivent les données, ou celles extraites à partir de ces données sont donc très bénéfiques pour la configuration automatique du processus d'interconnexion. Le résultat est, dans la plupart des cas, un ensemble de règles de spécifications d'interconnexion qui décrivent globalement le processus d'interconnexion. Il s'agit principalement de règles de sélection des ressources à interconnecter et des propriétés à comparer pour le faire. Ces techniques restent cependant globales et ne permettent donc pas d'adapter la configuration du processus d'appariement à un niveau plus fin. Par exemple, elles permettent de choisir les différents critères de comparaison, mais ne permettent pas forcément de choisir le poids de chaque critère. Nous verrons dans la suite les approches qui tentent d'adapter encore plus la configuration du processus d'interconnexion à un niveau plus fin.

Adaptation fine du processus d'interconnexion aux données

Extraire ou apprendre automatiquement les spécifications d'interconnexion peut faciliter le travail de l'utilisateur de l'outil d'interconnexion en réduisant son intervention. Cependant, nous considérons que dans le cas où les données présentent une forte hétérogénéité au sein d'une même source de données, la configuration d'interconnexion nécessite d'être adaptée aux spécificités de chaque ressource.

L'adaptation du paramétrage du processus d'interconnexion a été proposée pour des finalités différentes. Par exemple, (Seddiqui et al., 2015) propose une approche de génération automatique des poids pour chaque critère d'interconnexion dans un processus à stratégie multicritère. L'intuition sur laquelle repose cette approche est issue de la théorie de l'information où les valeurs les moins fréquentes sont considérées comme plus informatives que les valeurs les plus fréquentes. La technique utilisée est basée sur la combinaison de deux rapports calculés pour chaque propriété à partir d'un ensemble d'apprentissage : le rapport du nombre des valeurs distinctes de la propriété au nombre des ressources pour lesquelles cette propriété est instanciée, ainsi que le rapport du nombre des valeurs distinctes de la propriété au nombre total des ressources. L'adaptation du paramétrage se fait d'une manière statique en amont de l'exécution de la phase de mise en correspondance. L'outil **OKKAM** propose une technique un peu plus dynamique pour affecter des poids aux différents critères dans une comparaison multicritère. Dans cet outil, le score d'une mesure de comparaison des valeurs de deux propriétés est en fait pondéré selon la similarité de ces propriétés. (Georgala, 2016) propose une approche d'adaptation dynamique des spécifications du processus d'interconnexion à des fins d'optimisation de temps d'exécution. L'idée est d'adapter dynamiquement les spécifications d'interconnexion, *c.-à-d.* apprendre à partir des liens déjà découverts, pendant l'exécution, un ensemble de spécifications d'interconnexion inclus dans les spécifications initiales, mais qui garantit un temps d'exécution moins important, tout en gardant le même rappel.

Bien que les approches ci-dessus proposent des techniques d'adaptation du paramétrage d'interconnexion, elles ne répondent pas au besoin d'adaptation de la comparaison des géométries à leurs hétérogénéités. L'approche d'interconnexion proposée par (Salas et Harth, 2011), qui est basée sur la comparaison de géométries surfaciques par une distance de *Hausdorff*, se distingue dans ce contexte. Cette approche propose une technique d'adaptation du seuil d'appariement d'une manière dynamique. Cette approche estime, comme hypothèse de base, que le seuil de distance entre deux géométries dépend des surfaces des géométries en question. Tout d'abord, un échantillon

représentatif extrait du jeu de données source. Ensuite, une valeur de distance « moyenne » est calculée pour chaque géométrie de cet échantillon. Une valeur de distance « moyenne » représente pour chaque géométrie la moyenne entre la distance à la plus proche géométrie dans le jeu de données cible (*c.-à-d.* celle de son homologue) et la distance à la deuxième plus proche géométrie (représentant un objet non-homologue). Une fonction de seuil est calculée par une régression quadratique des valeurs de distances « moyennes » par rapport aux surfaces des géométries. La fonction de seuil apprise permet donc d'affecter un seuil de distance adapté à la taille de la géométrie du jeu de données source tout au long de l'exécution de la phase de mise en correspondance. La technique d'adaptation utilisée dans cette approche montre le grand potentiel de la prise en compte des spécificités de chaque géométrie dans le processus d'interconnexion.

2.3 Conclusion de l'état de l'art

Dans ce chapitre nous avons exposé les différentes tendances des approches de l'état de l'art dans deux contextes distincts, mais dont l'objectif demeure d'identifier des relations de correspondance entre les données. Les approches d'appariement de données géographiques, tout comme celles d'interconnexion de données sur le Web, ont pour objectif principal de mettre en correspondance des objets (des ressources) homologues ou de trouver un autre type de relations entre eux (elles). L'hétérogénéité des données une difficulté majeure dans la conception des approches des deux domaines. Une partie importante des approches d'appariement de données géographiques sont conçues en prenant en compte des a priori sur les données : le type du phénomène géographique représenté façonne la manière dont les données sont traitées par les algorithmes d'appariement. Par exemple, s'il s'agit de données représentant de réseaux (*ex.* routier ou hydrographique) certains algorithmes sont conçus en supposant que les géométries présentent une organisation topologique (*ex.* (Mustière et Devogele, 2008) et (Volz, 2006)) ou hiérarchique (*ex.* (Cost, 2014)) et s'en servent pour l'appariement. S'il s'agit de phénomènes avec une emprise au sol importante (*ex.* les bâtiments, les cultures, les forêts) l'algorithme suppose un recouvrement important entre les géométries et s'en sert pour l'appariement, comme c'est le cas dans l'approche de (Bel Hadj Ali, 2001). Les différences de précisions géométriques ou de niveaux de détail sont souvent prises en compte par les algorithmes d'appariement via le paramétrage du processus ou le prétraitement des données. Par exemple, pour gérer la différence de précision géométrique, l'algorithme peut s'appuyer sur un recalage des géométries et une estimation du seuil de distance à partir des données (*ex.* (Walter et Fritch, 1999)). Pour remédier aux différences de niveau de détail l'algorithme peut avoir recours à une homogénéisation du niveau de détail entre les deux jeux de données comme c'est le cas pour (Yang et al., 2014). Dans ces exemples, les solutions apportées pour résoudre les hétérogénéités sont basées sur des a priori que le concepteur de l'algorithme d'appariement a sur les données. Les connaissances qu'il possède sur les données sont à chaque fois codées en dur dans l'algorithme proposé. Ceci constitue la différence principale avec les approches sans a priori sur les données qui laissent à l'utilisateur de l'algorithme le soin de spécifier ces connaissances sur les données comme c'est le cas de (Olteanu, 2008), ou s'appuie sur des connaissances externes telles l'alignement des schémas dans des approches comme (Resseler et al., 2009) ou (Uitermark et al., 2001). Les propositions de (Gesbert, 2005) et (Abadie, 2012) se distinguent notamment par leur prise en compte des spécifications des bases de données géographiques dans l'alignement de leurs schémas. Même si la formalisation des spécifications dans ce cas a pour but de chercher des alignements entre schémas et non pas un appariement entre les objets géographiques en soi, ces travaux représentent une piste importante pour la prise en compte automatique de la différence des représentations

géométriques dans la comparaison des géométries. Les approches d'interconnexion des données du Web se distinguent particulièrement par leur généricité et leur prise en compte des connaissances formalisées, apprises à partir des données ou provenant d'une source externe, pour améliorer le processus d'interconnexion, notamment pour automatiser sa configuration. **Silk** et **LIMES** demeurent les outils les plus génériques qui incluent des possibilités de configuration manuelle ou automatique par apprentissage, des méthodes d'optimisation de la complexité temporelle et une prise en compte de l'information géométrique en proposant un ensemble de mesures de distance spatiale et de calcul de relations topologiques.

Conclusion de la partie A

Les références spatiales des ressources publiées sur le Web de données constituent une information qui peut être très bénéfique pour l'interconnexion de ces ressources. En effet, celles-ci peuvent être vues comme des clés (ou des quasi-clés) permettant de décider de la mise en correspondance de ressources. Cependant, l'expérience acquise dans le domaine de l'appariement de données géographiques montre qu'elles peuvent présenter différents types d'hétérogénéités, généralement liés à leur niveau de détail, qui rendent leur utilisation à des fins de mise en correspondance complexe. De plus, ces références spatiales peuvent être mises à profit pour produire des applications de visualisation cartographique des données. Celles-ci permettent d'explorer les données de façon intuitive voire d'accéder à des connaissances sur les données à un plus haut niveau d'abstraction (*ex.* identifier leur emprise spatiale, détecter des phénomènes de corrélation spatiale, etc.). Dans ce cas de nouveau, le niveau de détail des données et leur échelle de visualisation vont jouer un rôle important dans l'efficacité des solutions de visualisation proposées.

Les approches d'appariement de données géographique proposent des solutions astucieuses pour résoudre les hétérogénéités, mais qui s'appliquent à des types de données et d'hétérogénéités bien spécifiques. Les implémentations de ces approches d'appariement manquent souvent de généricité et sont rarement utilisables pour d'autres cas d'application que celui pour lequel elles ont été conçues. Les bases de données géographiques traitées dans les approches d'appariement sont caractérisées par une homogénéité interne qui permet de traiter leurs objets d'une manière similaire pendant tout le processus d'appariement. Sur le Web de données, où les références spatiales proviennent de sources multiples, ces hétérogénéités peuvent se manifester au sein d'une même source de données (*cf.* section 1.4.2).

Par rapport aux approches existantes d'appariement de données géographiques, nous souhaitons principalement améliorer deux aspects: la généricité et le niveau de granularité de la prise en compte des hétérogénéités des géométries dans l'interconnexion.

L'état de l'art sur les approches et outils d'interconnexion de données liées comprend une variété d'approches qui s'appuient sur les connaissances qui décrivent les données ou des connaissances externes afin de paramétrer le processus d'interconnexion. Bien que cela constitue une piste intéressante pour la prise en compte des hétérogénéités dans la configuration d'un processus d'interconnexion, ces approches restent très limitées vis-à-vis des problèmes d'hétérogénéité intra-sources de données. La configuration (manuelle ou automatique) d'un processus d'interconnexion de données du Web demeure la même tout au long de la phase de mise en correspondance, *c.-à-d.* toutes les ressources sont comparées de la même manière. Peu d'approches proposent d'adapter la comparaison des ressources en prenant compte les différences de leurs caractéristiques à l'intérieur d'une même source. L'approche proposée par (Salas et Harth, 2011) illustre ce cas précisément, en fournissant pour chaque paire de géométries comparées, un seuil adapté à leurs valeurs de superficie. L'utilisation de règles de décision pour formaliser l'expertise de l'utilisateur dans les choix de configuration de l'alignement dans l'approche proposée par (Mochol et Jentzch, 2008) constitue une piste intéressante également.

Le problème principal qui se pose est : comment peut-on configurer un processus d'interconnexion générique basé sur la comparaison des géométries qui prend en compte les hétérogénéités

géométriques inter et intra sources de données ? Celle-ci requiert des connaissances a priori sur les caractéristiques des jeux de données en question pour fixer des paramètres tels que le choix de mesure de distance, le seuil de comparaison géométrique ou le poids de cette comparaison dans le cas d'un processus multicritère.

Un second frein à la bonne utilisation des références spatiales et des liens d'interconnexion entre données que nous tâcherons de résoudre est celui de leur visualisation cartographique à différentes échelles : comment proposer une interface de visualisation cartographique générique au sein de laquelle l'information reste lisible à différentes échelles.

PARTIE B
PROPOSITIONS

Propositions pour une meilleure prise en compte du niveau de détail et de la modalisation géométrique des ressources géoréférencées sur le Web de données.

Introduction

Dans la partie A de ce mémoire, nous avons vu comment les questions liées au niveau de détail et à la modélisation géométrique des références spatiales associées aux ressources du Web de données peuvent causer des difficultés pour la mise en correspondance ou la visualisation cartographique de ces ressources. Si des solutions ont été proposées dans le domaine des sciences de l'information géographique, leur transposition directe sur les ressources du Web de données laisse à désirer. En effet, les ressources géoréférencées du Web de données présentent des caractéristiques différentes des bases de données géographiques et nécessitent un traitement dédié. Dans cette partie, nous tentons de proposer des solutions en nous appuyant à la fois sur l'état de l'art en science de l'information géographique et sur les possibilités nouvelles offertes par les standards du Web de données.

Pour remédier aux difficultés de mise en œuvre des références spatiales comme critère de mise en correspondance, nous proposons de revenir aux causes mêmes des hétérogénéités géométriques. Ceci revient à répondre à la question : quelles sont les caractéristiques des géométries dont la différence constitue une hétérogénéité qui peut affecter la comparaison des géométries dans un processus d'interconnexion ? Notre hypothèse de base est que la prise en compte de ces caractéristiques dans un processus d'interconnexion qui utilise la comparaison des géométries comme critère peut améliorer la qualité de ses résultats. Nous considérons ces caractéristiques comme des métadonnées sur les géométries. Nous proposons d'inclure ces métadonnées comme des connaissances supplémentaires dans le processus d'interconnexion à l'instar des approches existantes basées sur des connaissances internes ou externes. Pour cela, nous proposons de représenter formellement ces connaissances pour qu'elles puissent être chargées et exploitées avec les données dans le processus d'interconnexion.

Pour améliorer la lisibilité des ressources du Web affichées sur des interfaces cartographiques, nous proposons de modifier leur niveau de détail en fonction de l'échelle d'affichage. Ceci peut être réalisé en exploitant les liens d'interconnexion entre ces ressources et celles d'un référentiel de données géographiques. Lorsque les géométries de référence sont plus détaillées, elles peuvent servir à afficher les données à plus grande échelle. Dans le cas contraire, on peut procéder à une agrégation pour afficher les données de façon lisible à plus petite échelle. Pour réaliser de telles opérations, nous proposons de nous appuyer sur les vocabulaires utilisés pour décrire les ressources du Web.

3 FORMALISATION ET ACQUISITION DES CONNAISSANCES POUR LA QUALIFICATION DES RÉFÉRENCES SPATIALES DIRECTES SUR LE web DE DONNÉES

Le paramétrage d'un processus d'interconnexion fondé sur la comparaison de références spatiales nécessite de disposer de connaissance sur les caractéristiques de ces références spatiales. En effet, des caractéristiques différentes d'une géométrie à une autre engendrent des hétérogénéités. Les caractéristiques des géométries doivent donc être formalisées pour pouvoir être prises en compte automatiquement dans un processus d'interconnexion. Représenter formellement ces caractéristiques sous forme de connaissances exploitables nécessite d'abord la définition d'un vocabulaire. Nous présentons dans cette partie les choix de modélisation de ce vocabulaire ainsi qu'une approche d'acquisition des connaissances pour associer à chaque géométrie d'un jeu de données des métadonnées sur ses caractéristiques formalisées conformément à notre vocabulaire.

3.1 Un vocabulaire pour décrire la sémantique des XY

Dans la partie 1.1.3 nous avons identifié différentes causes d'hétérogénéité géométrique inter et intra sources de données. Nous avons vu que différents points de vue sur le monde réel peuvent se manifester par des niveaux de détail différents d'un jeu de données à un autre. Ceci se concrétise par des spécifications et des processus de saisie différents et donc des représentations géométriques différentes d'une source de données à une autre. En outre, les erreurs de saisie liées aux facteurs humain et matériel, ainsi que la nature ouverte, souvent collaborative, de certaines sources de données peuvent accentuer les hétérogénéités géométriques au sein d'un même jeu de données. Un niveau de détail géométrique et des spécifications de données bien définis nous permettent de comprendre le sens de chaque géométrie : quel type d'entité géographique elle représente, quelles sont ses règles de saisie, quel est l'élément caractéristique de la forme des entités géographique choisi pour sa modélisation, quelle est sa précision, etc. En d'autres termes, ils nous permettent de comprendre quelle est la **sémantique** portée par chaque géométrie. Les hétérogénéités entre les géométries ne sont donc rien d'autre que des différences de sémantique entre ces dernières. Nous appelons donc **sémantique des XY** l'ensemble des caractéristiques d'une géométrie liées à son niveau de détail et à ses spécifications de saisie, qui permettent de comprendre le sens de la géométrie et dont les différences d'une ressource à une autre engendrent des hétérogénéités géométriques. À partir des éléments évoqués dans la partie 1.1.3, ainsi que les hétérogénéités prises en compte par les approches d'appariement de l'état de l'art présentées dans la partie 2.1, nous avons identifié quatre caractéristiques des géométries principales permettant de qualifier une géométrie:

- Sa précision planimétrique
- Sa modélisation géométrique
- Le caractère plus ou moins vague de l'entité géographique qu'elle représente
- Sa résolution géométrique

Nous proposons un vocabulaire de la sémantique des XY qui permet d'expliciter ces caractéristiques et donc d'exploiter ces connaissances pour paramétrer automatiquement un processus d'interconnexion. Nous nous arrêtons en priorité sur ces quatre caractéristiques pour deux raisons. D'une part, elles sont les plus importantes pour identifier et comprendre les hétérogénéités entre

géométrie. D'autre part, contrairement à d'autres caractéristiques des géométries telles que l'orientation, l'élongation, l'aire, etc. qui sont implicitement présentes dans la géométrie et qui ne sont pas difficiles à extraire à la volée, ces caractéristiques nécessitent soit une très bonne connaissance des processus d'acquisition des données, soit des analyses élaborées pour être connues.

Les connaissances à propos des caractéristiques des géométries en pouvant être à l'origine d'hétérogénéités peuvent être perçues comme des métadonnées sur ces géométries. De nombreuses ontologies sont dédiées à la représentation des métadonnées des données publiées sur le Web, telles que **DCE**[67] (*Dublin Core Metadata Element Set*), **DCAT**[68] (*Data Catalog Vocabulary*), **VoID**[69], **PROV-O**[70], etc. Ces vocabulaires visent principalement à réduire l'écart entre le fournisseur et l'utilisateur des données. Le vocabulaire **DCE** fait partie de l'initiative « *DCMI Metadata Terms* »[71] qui assure le maintien de nombreux vocabulaires et spécifications liées à la représentation des métadonnées. Le vocabulaire **DCE** inclut une quinzaine de propriétés utilisées dans la description des ressources, *ex.* le format, les dates des différents évènements de son cycle de vie (création, modification, etc.), la langue, la source, le titre, etc. **DCAT** est un vocabulaire recommandé par le W3C pour faciliter l'interopérabilité des catalogues de données sur le Web. Ce vocabulaire réutilise des propriétés d'autres vocabulaires pour définir un modèle de description et de structuration des métadonnées des jeux de données dans des catalogues. **VoID** est un autre vocabulaire de description des métadonnées qui s'intéresse également aux aspects de découverte et de catalogage des jeux de données. Il se distingue notamment par les possibilités qu'il offre pour la description des métadonnées sur la structure des données, sur les méthodes d'accès aux données, ainsi que sur les liens avec d'autres jeux de données. Le vocabulaire **PROV-O** sert à représenter et échanger les métadonnées de provenance des données. Il comprend trois classes principales (entité, activité et agent) ainsi que les propriétés qui les relient. Ce vocabulaire peut, sur cette base, être utilisé directement ou être étendu en vocabulaires spécifiques aux différents domaines d'application. Les vocabulaires génériques pour décrire les métadonnées ne semblent pas suffisants pour répondre à notre besoin de représentation des caractéristiques des géométries. Nous proposons donc un nouveau vocabulaire qui s'inspire plutôt des modèles de métadonnées dédiés aux données géographiques.

Ce vocabulaire doit à la fois tenir compte des modes de représentation des références spatiales des ressources sur le Web de données et des problématiques d'hétérogénéités géométriques posées dans le cadre d'un processus d'interconnexion. Tout d'abord, nous devons choisir la portée des métadonnées géométriques, *c.-à-d.* le niveau auquel elles devront être représentées : au niveau du jeu de données, au niveau de chaque ressource géoréférencée ou au niveau des géométries. Représenter ces métadonnées comme des métadonnées générales d'un jeu de données suggère qu'elles sont les mêmes pour toute géométrie de chaque ressource présente dans ce jeu. Or, nous souhaitons interconnecter des jeux de données présentant des hétérogénéités géométriques internes. Représenter ces métadonnées au niveau de chaque ressource géoréférencée semble un

[67] http://purl.org/dc/elements/1.1/
[68] https://www.w3.org/TR/vocab-dcat/
[69] https://www.w3.org/TR/void/
[70] https://www.w3.org/TR/prov-o/
[71] http://dublincore.org/documents/dcmi-terms/

choix plus adéquat. Cependant, certaines ressources peuvent être liées à plusieurs géométries, présentant chacune des caractéristiques différentes (voir Figure 3.1). Nous proposons donc de lier ces métadonnées directement à chaque géométrie (c.-à-d. à chaque ressource de type « Géométrie » associé qui représente la géométrie).

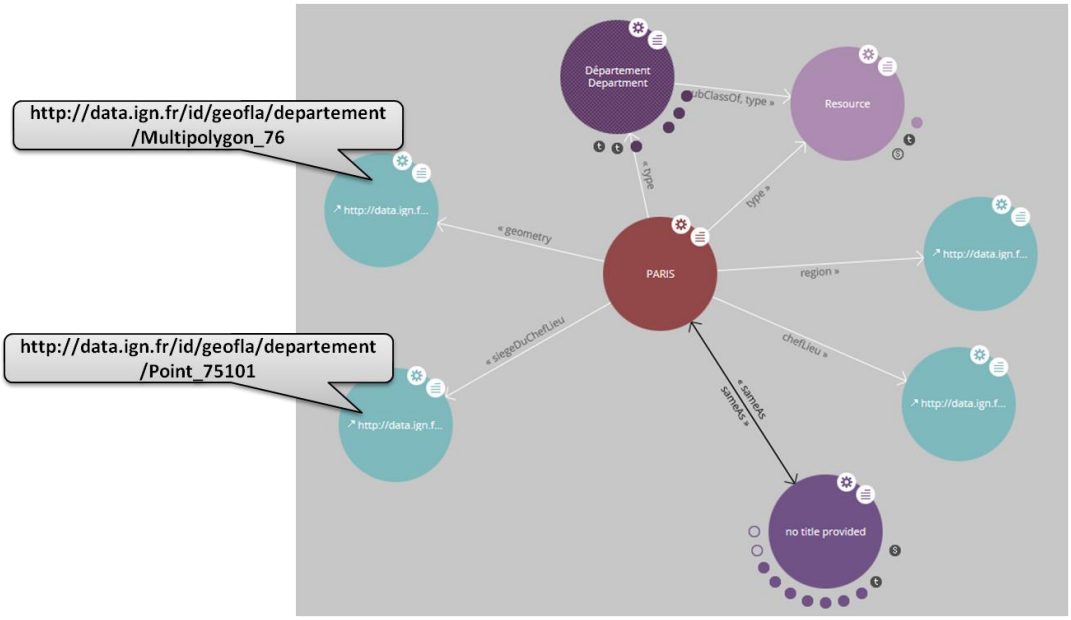


Figure 3.1 Exemple de ressource ayant plusieurs géométries : la ressource http://data.ign.fr/id/geofla/departement/75 représentant le département de Paris dans le jeu de données Geofla, la première géométrie représente la limite du département, la deuxième représente la localisation du siège de son chef-lieu (visualisation réalisées avec http://en.lodlive.it)

En l'absence de standard encore bien installé pour la représentation des géométries sur le Web de données, nous proposons d'étendre l'ontologie des géométries geom[72] (*cf.* section 1.1.2) qui présente des avantages tels que la compatibilité avec les ontologies GeoSPARQL et NeoGeo ainsi que la possibilité de représenter des géométries de façon structurée (Hamdi et al., 2014). Nous présentons dans la suite nos choix de modélisation pour chacune des caractéristiques des géométries que nous avons choisi de décrire et de mettre à profit pour paramétrer un processus d'interconnexion.

3.1.1 Métadonnée sur la précision planimétrique des géométries

État de l'art sur la représentation des métadonnées sur la précision planimétrique

La qualité de données dans un sens général est définie par la norme ISO 8402 comme étant « l'ensemble des propriétés et caractéristiques d'un produit ou d'un service qui lui confère l'aptitude à satisfaire des besoins exprimés ou implicites ». Dans le domaine des bases de données géographiques, la précision planimétrique des géométries est un critère de qualité de données qui représente l'écart en planimétrie entre la position d'une géométrie dans une base de données géographiques et celle de l'entité réelle qu'elle représente.

[72] http://data.ign.fr/def/geometrie

Selon la norme ISO 19115-1 la qualité de données géographique est renseignée comme une métadonnée. La norme ISO 19157 définit la qualité des données géographiques comme étant la composition de plusieurs « éléments de qualité », qui réfèrent, comme toutes autres métadonnées, à un domaine d'application (ou portée). Selon cette norme, la précision planimétrique est, entre autres, une spécialisation du concept général d'« élément de qualité ». Un élément de qualité peut être spécialisé en plusieurs sous-éléments comme: l'exhaustivité, la précision thématique, la cohérence logique, l'utilisabilité, la qualité temporelle ou la précision de localisation des objets. Un élément de qualité est évalué par une « mesure de référence » en suivant «une méthode d'évaluation » qui fournit un « résultat ». Selon l'élément de qualité évalué, le « résultat » peut être : de « conformité », « descriptif » ou « quantitatif ». Dans le cas d'un élément de qualité représentant la précision planimétrique le résultat est quantitatif. Il est représenté par une valeur, le type de cette valeur et son unité de mesure. Pour obtenir ces résultats, les méthodes d'évaluation appliquées ne sont pas forcément directes ; un résultat peut être déduit par agrégation ou une dérivation d'autres résultats, comme il peut être indirectement déduit à partir d'une connaissance externe ou l'expérience que l'utilisateur a du produit. Les connaissances externes peuvent inclure toute information non quantitative, comme la généalogie ou l'objectif des données (décrits dans ISO 19115-1) ou tout autre rapport de qualité sur les données utilisées pour constituer le jeu de données. Les métadonnées doivent préciser, selon la norme ISO 19115-1 leur domaine d'application (*scope*), qui désigne par un code le niveau auquel la métadonnée est appliquée (jeu de données, collection*, feautre type*, *feature*, attribut …). Les spécifications de la directive INSPIRE précisent qu'il faut se référer à la norme ISO 19157 pour rapporter la qualité des données et choisir les mesures utilisées pour son évaluation, y compris la précision planimétrique des géométries des données. Ceci est présent dans tous les documents[73] de spécifications des différents thèmes de la directive INSPIRE au niveau des recommandations concernant la qualité des données. Dans la pratique, dans le cas des jeux de données géographiques, la précision planimétrique est généralement fournie dans des métadonnées séparées, dans des documents descriptifs de chaque jeu de données. Elle peut également être mentionnée sous forme d'un attribut au niveau de chaque instance des données. Les figures 3.2, 3.3 et 3.4 sont des exemples des précisions planimétriques renseignées dans les spécifications de contenu des bases de données géographique vectorielles produites par l'IGN : BD TOPO®, BD CARTO® et BD ADRESSE®.

Source des données	Précision	Traduction dans l'attribut « précision planimétrique » PREC_PLANI
Photogrammétrie, plan ou fichier métrique	0,5 à 1,5 m	1.5
Levé GPS dynamique, BD TOPO® version antérieure, BD PARCELLAIRE® recalée	1,5 à 2,5 m	2.5
Orthophotographie, plan ou fichier non métrique, levé terrain, BD PARCELLAIRE®	2,5 à 5 m	5
Carte 1/25000 (SCAN 25®), calculé, image satellite	5 m à 10 m	10
BD CARTO®, GEOROUTE®	> à 10 m	30

Figure 3.2 Précision planimétrique des données BD TOPO® et la BD ADRESSE® selon leur source. Extrait des spécifications de la base BD TOPO® 2.2 (http://professionnels.ign.fr/doc/DC-BDTOPO-2-2.pdf) et de la BD ADRESSE® 2.1 (http://professionnels.ign.fr/doc/DC_BDADRESSE_2-1.pdf).

[73] http://inspire.ec.europa.eu/index.cfm/pageid/2

Précision géométrique
Les mesures de la qualité géométrique font état d'une précision qui varie, selon les thèmes, entre 15 et 50 m en erreur moyenne quadratique.

Figure 3.3 Précision planimétrique des données BD CARTO®. Extrait des spécifications de la base BD CARTO®3.2 (http://professionnels.ign.fr/sites/default/files/DC_BDCARTO_3-2.pdf).

Projection du centre des parcelles	Moyenne des différences géométriques entre les coordonnées dans la base et les coordonnées terrain : **11 m** (écart type : **18 m)**
Plaque adresse	Moyenne des différences géométriques entre les coordonnées dans la base et les coordonnées terrain : **6 m** (écart type : **12 m)**

Figure 3.4 Précision planimétrique des données BD ADRESSE® selon leurs types de localisation. Extrait des spécifications de la base BD ADRESSE®2.1 (http://professionnels.ign.fr/doc/DC_BDADRESSE_2-1.pdf).

Sur le Web de données, ou d'une manière plus générale dans les données ouvertes sur le Web, on trouve rarement des indications sur la précision planimétrique des données géoréférencées. Quand elle est fournie, la précision planimétrique peut être présente en tant que métadonnée générale d'un jeu de données. La base des adresses du Grand Lyon[74], ou la base des points de relevés d'analyse d'eau de l'ONEMA[75] sont deux exemples de bases de données dont la précision planimétrique est fournie comme une métadonnée.

Tout comme le cas de l'estimation indirecte de la qualité des données géographiques, la généalogie des données permet dans certains cas d'estimer la précision des références spatiales. Dans les jeux de données du Web l'information de généalogie peut être fournie sous forme de description textuelle de l'historique des données, sur moyens, les procédures et les traitements mis en œuvre ou sur l'entité (personne ou organisme) en charge pour la production de ces données. Le champ « Producteur » dans la base de données base des points de relevés d'analyse d'eau de l'ONEMA[76], ou le champ « TypCoordLieuSurv » de la base des lieux de surveillance des eaux littorales produite par le SANDRE[77].

Un travail de représentation de la norme ISO 19115 sous la forme d'ontologies OWL a été réalisé par (Cox, 2012). Le même auteur a proposé une ontologie **dq**[78] regroupant la partie des métadonnées qui constitue le noyau du modèle de la qualité de données décrit dans la norme ISO 19157. D'autres travaux ont traité de manière plus générale la question de la qualité de données, et proposent des ontologies pour la représentation de la qualité des données. Par exemple l'ontologie **daQ**[79] proposée

[74] https://data.grandlyon.com/localisation/voies-et-adresses-du-grand-lyon/
[75] https://www.data.gouv.fr/fr/datasets/points-de-prelevements-associes-aux-stations-de-la-mesure-de-la-qualite-des-eaux-superficielles--onm/
[76] https://www.data.gouv.fr/fr/datasets/station-de-mesure-de-la-qualite-des-eaux-superficielles-continentales-metropole/
[77] http://www.data.eaufrance.fr/jdd/a0a78ff9-f3fe-4211-a6d8-43dc82940aa7
[78] http://def.seegrid.csiro.au/isotc211/iso19115/2003/dataquality
[79] http://purl.org/eis/vocab/daq

par (Debattista et al., 2014) permet de stocker, dans un graphe séparé et lié à un graphe de données, des informations sur une qualité observée en précisant la « métrique » utilisée et la « valeur » calculée pour une ressource. Chaque métrique est liée à une « dimension » spécifique, *c.-à-d.* à un aspect de la qualité des données que l'on cherche à évaluer. Selon les auteurs ce modèle peut être étendu pour comprendre des dimensions et des métriques bien spécifiques au cas d'application. Cette ontologie peut donc être étendue pour l'appliquer aux données géoréférencées. Dans ce cas, les différentes classes d'éléments de qualité dans le modèle de la norme ISO 19157 peuvent correspondre aux différentes « dimensions » de l'ontologie **daQ**.

Représentation de la précision planimétrique dans le vocabulaire de la sémantique des XY

Notre vocabulaire doit permettre d'associer à chaque géométrie un élément de précision planimétrique qui est décrit par une méthode et un résultat d'évaluation. La méthode d'évaluation permet de préciser s'il s'agit d'une dérivation à partir d'un autre élément de qualité ou d'une évaluation à partir des données. C'est ce deuxième cas qui le plus souvent employé. Une évaluation à partir des données peut être réalisée directement par inspection complète ou par analyse d'un échantillon des données, ou indirectement à partir d'une connaissance sur la généalogie des données. Le résultat d'évaluation est décrit par sa valeur numérique ainsi que son unité de mesure.

Bien que l'ontologie **daQ** permette d'exprimer le résultat d'évaluation de la précision planimétrique, elle ne fournit pas de moyen pour indiquer le type de méthode d'évaluation utilisé. Or, nous considérons cette information comme primordial à l'interprétation de la valeur de précision planimétrique. L'ontologie **PROV-O** permet de décrire des métadonnées de généalogie. Elle peut être utile pour présenter la source de déduction d'une évaluation indirecte, mais ne suffit pas pour exprimer le résultat d'évaluation. Nous avons donc décidé de nous appuyer principalement sur le noyau de la norme ISO 19157 pour la représentation de la précision planimétrique (voir Figure 3.5). L'ontologie **dq** proposée par (Cox, 2012) pour la représentation des éléments de cette norme ISO 19157 demeure compliquée à utiliser directement pour le simple cas de la représentation de la précision planimétrique au niveau de chaque géométrie : il n'existe pas de propriété qui permet d'associer à chaque géométrie sa précision planimétrique. Nous proposons donc notre propre vocabulaire qui s'appuie sur la norme ISO 19157 tout en faisant référence aux éléments correspondants (classes et propriétés) dans **dq**, à l'aide de propriétés owl:equivalentClass. Pour les unités de mesures, nous réalisons directement le vocabulaire QUDT[80].

La Figure 3.6 présente un exemple de représentation en RDF de la précision planimétrique d'une géométrie.

[80] http://qudt.org/schema/qudt

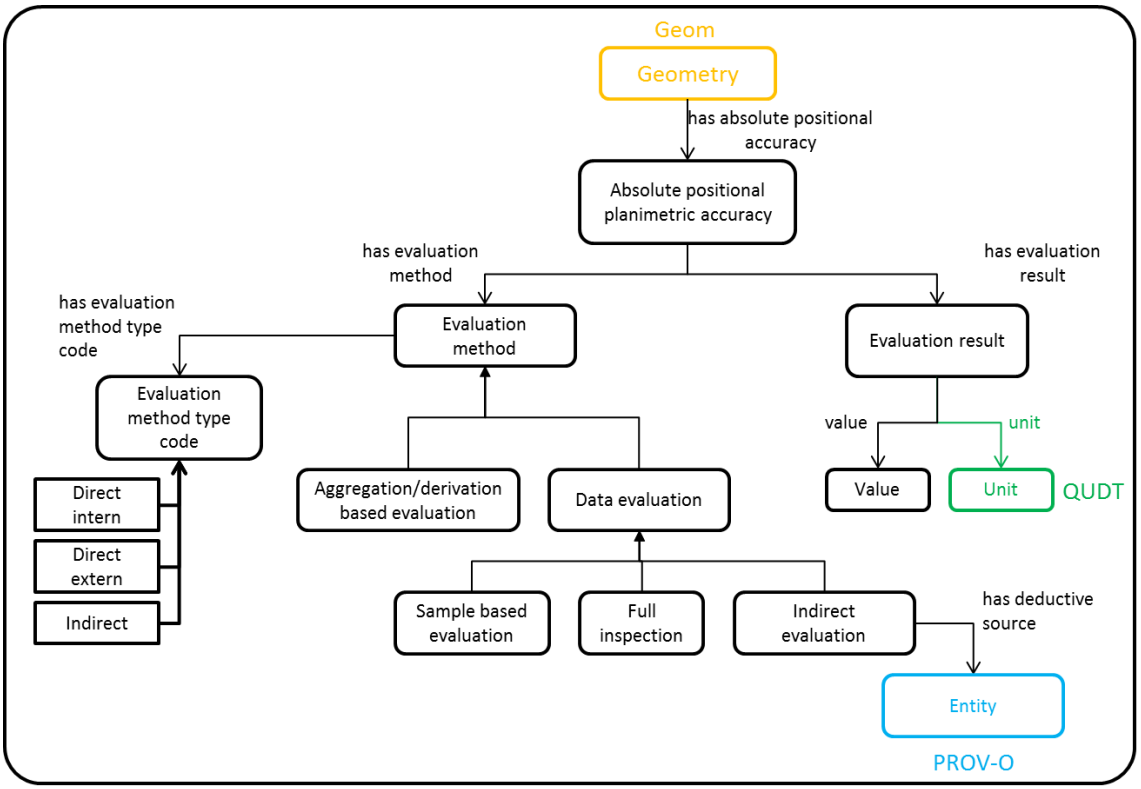


Figure 3.5 Extrait du vocabulaire de la sémantique des XY pour représenter la précision planimétrique des géométries.

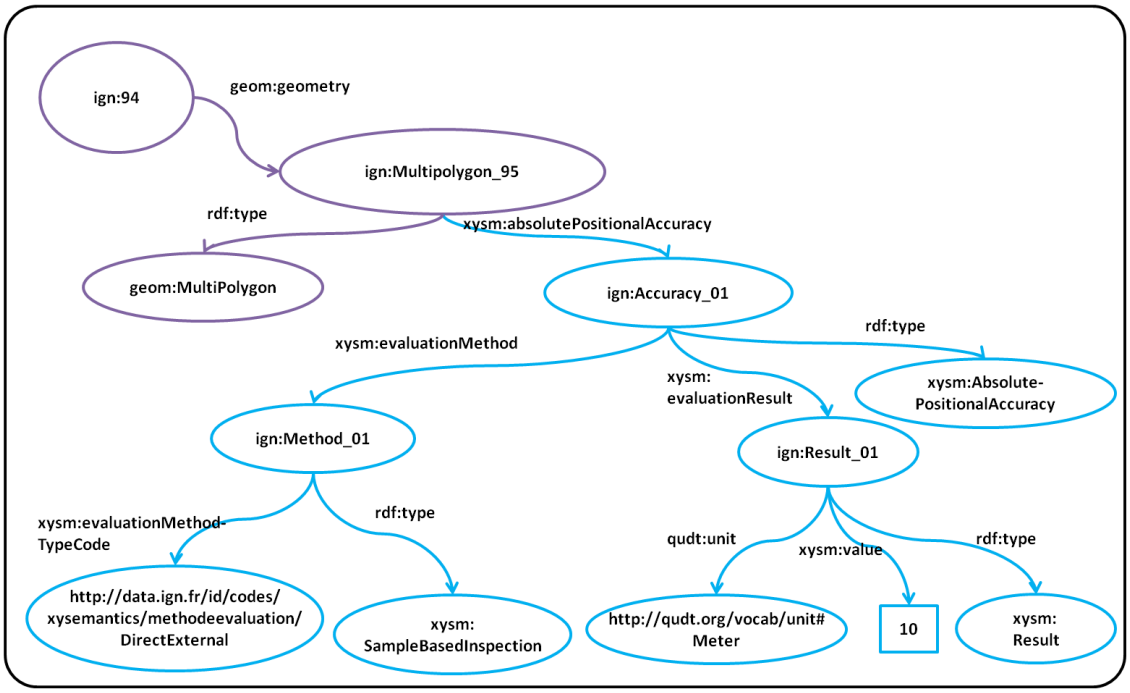


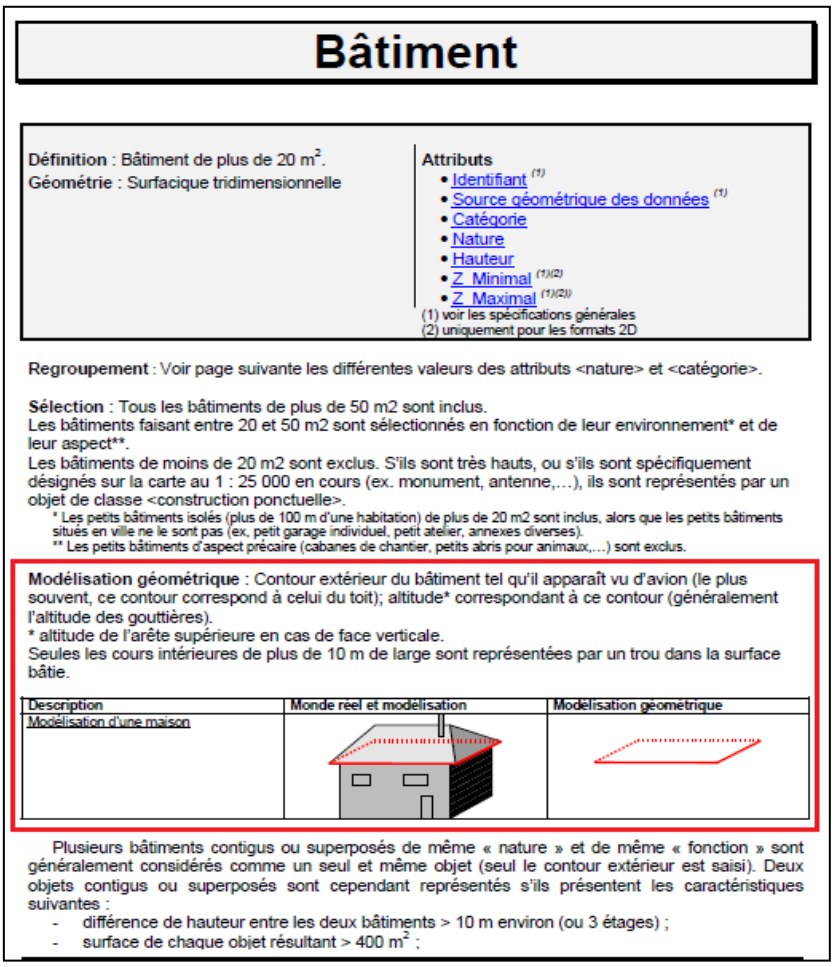
Figure 3.6 Représentation en RDF de la précision planimétrique de la géométrie délimitant le département 94 dans la source Geofla ign[81]:94

[81] http://data.ign.fr/id/geofla/departement/

3.1.2 Métadonnée sur la modélisation géométrique

Etat de l'art sur la représentation des métadonnées sur la modélisation géométrique

Dans une base de données géographique, la modélisation géométrique représente les choix de représentation géométrique effectués par rapport à l'entité du monde réel représentée par un objet géographique. La modélisation géométrique est une information centrale dans les spécifications de saisie des données géographiques. Elle consiste à préciser quelle primitive géométrique doit être utilisée pour représenter chaque type d'entité géographique ainsi que les éléments qui caractérisent spatialement les entités géographiques du monde réel (comme le contour d'un bâtiment ou l'axe d'une route) qui seront utilisés comme repères pour saisir la géométrie. La norme ISO 19131, qui a pour objectif de permettre la documentation des spécifications de saisie des bases de données géographiques, comprend une partie qui sert à décrire la modélisation géométrique ainsi que les règles de saisie des données. Cette partie nommée « *data capture* » n'est autre qu'un champ textuel contenant un paragraphe libre. Les spécifications du contenu des bases de données géographiques de l'IGN, par exemple, contiennent pour chaque classe une description de sa modélisation géométrique (voir Figure 3.7). Les fiches descriptives des produits de l'IGN contiennent également, entre autres informations, une description textuelle et illustrée de la modélisation géométrique des instances de chaque classe (*ex.* voir Figure 3.8).



Bâtiment

Définition : Bâtiment de plus de 20 m².
Géométrie : Surfacique tridimensionnelle

Attributs
- Identifiant (1)
- Source géométrique des données (1)
- Catégorie
- Nature
- Hauteur
- Z_Minimal (1)(2)
- Z_Maximal (1)(2)

(1) voir les spécifications générales
(2) uniquement pour les formats 2D

Regroupement : Voir page suivante les différentes valeurs des attributs <nature> et <catégorie>.

Sélection : Tous les bâtiments de plus de 50 m2 sont inclus.
Les bâtiments faisant entre 20 et 50 m2 sont sélectionnés en fonction de leur environnement* et de leur aspect**.
Les bâtiments de moins de 20 m2 sont exclus. S'ils sont très hauts, ou s'ils sont spécifiquement désignés sur la carte au 1 : 25 000 en cours (ex. monument, antenne,…), ils sont représentés par un objet de classe <construction ponctuelle>.
* Les petits bâtiments isolés (plus de 100 m d'une habitation) de plus de 20 m2 sont inclus, alors que les petits bâtiments situés en ville ne le sont pas (ex, petit garage individuel, petit atelier, annexes diverses).
** Les petits bâtiments d'aspect précaire (cabanes de chantier, petits abris pour animaux,…) sont exclus.

Modélisation géométrique : Contour extérieur du bâtiment tel qu'il apparaît vu d'avion (le plus souvent, ce contour correspond à celui du toit); altitude* correspondant à ce contour (généralement l'altitude des gouttières).
* altitude de l'arête supérieure en cas de face verticale.
Seules les cours intérieures de plus de 10 m de large sont représentées par un trou dans la surface bâtie.

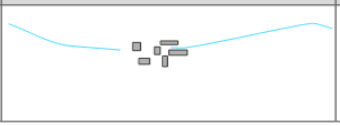
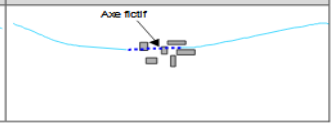
Description	Monde réel et modélisation	Modélisation géométrique
Modélisation d'une maison		

Plusieurs bâtiments contigus ou superposés de même « nature » et de même « fonction » sont généralement considérés comme un seul et même objet (seul le contour extérieur est saisi). Deux objets contigus ou superposés sont cependant représentés s'ils présentent les caractéristiques suivantes :
- différence de hauteur entre les deux bâtiments > 10 m environ (ou 3 étages) ;
- surface de chaque objet résultant > 400 m² ;

Figure 3.7 Extrait des spécifications de la classe BATIMENT de la base de données BD TOPO® 1.2

Modélisation géométrique :
À l'axe et à la surface du cours d'eau (tel qu'il se présente sur les photographies aériennes).

L'orientation de l'objet définit le sens d'écoulement. Elle n'est pas significative dans les zones très plates (ex : marais) ni pour les canaux.

Description	Monde réel	Modélisation géométrique
La continuité du réseau hydrographique est assurée par des éléments linéaires qui peuvent prendre la valeur d'attribut souterrain = oui (canal navigable) ou fictif = oui (autres cours d'eau)		

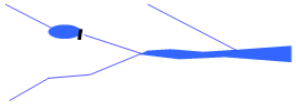
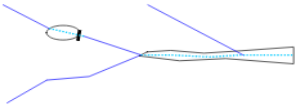
Description	Monde réel	Modélisation géométrique
Les éléments surfaciques sont doublés d'un objet <tronçon de cours d'eau> d'attribut <fictif> = « oui ».		

Figure 3.8 Extrait des spécifications de la classe TRONCON_COURS_EAU de la base de données BD TOPO® 2.2 (http://professionnels.ign.fr/sites/default/files/DC_BDTOPO_2-2.pdf)

Sur le Web également, certains jeux de données sont fournis avec des métadonnées qui décrivent leur contenu, souvent sous forme de fiches descriptives. C'est le cas, par exemple, des différents jeux de données d'adresses du Grand Lyon (voir Figure 3.9). Trois jeux de données de points d'adresses sont fournis avec chacun un choix de modélisation géométrique différent : point implanté sur le bâtiment, point implanté sur l'axe de la voie correspondante et point implanté à la position exacte du numéro (de la plaque) d'adresse. Chacun de ces choix de modélisation géométrique est décrit dans une fiche de métadonnées présentée en Figure 3.9.

Description sommaire

Les données se répartissent selon les couches suivantes:
Tronçon de trame viaire : Filaire à l'axe de la voie.
Nœud de trame viaire : intersection de tronçons.
Lieudit: découpage cadastral sur la commune.
Numéro de voirie : numéro implanté sur voie ou bâtiment.
Point de débouché : implanté sur le filaire au droit du n°.
Point d'adressage : point implanté sur bâtiments.
Nom de voie : toponymie
Nom de lieudit : toponymie du lieudit.
Dénomination : donnée alphanumérique liée au FUV

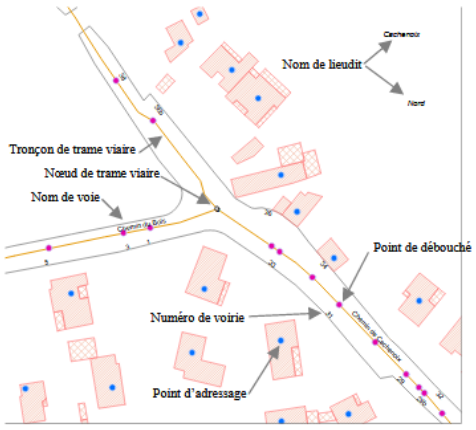


Figure 3.9 Extrait de la fiche[82] de métadonnées descriptives des jeux de données d'adressage du Grand Lyon.

Bien que les métadonnées sur la modalisation géométrique soient importantes pour décider de l'utilité des données pour une application et apprendre comment les utiliser, elles ne sont pas toujours fournies. Dans le cas où elles sont fournies avec les données, elles restent restreintes à des descriptions textuelles ou des illustrations. La formalisation des spécifications des bases de données géographiques a suscité l'intérêt de quelques travaux seulement. Dans un cadre d'intégration de données, (Uitermark, 2001) propose une approche qui s'appuie sur certaines connaissances sur les

[82] https://download.data.grandlyon.com/catalogue/srv/metadata/c373e341-06e5-48e2-a7f3-2729ff5e1399

spécifications des données géographiques (leurs règles de sélection) pour l'alignement des schémas ainsi que pour filtrer les résultats d'appariements des données. Dans une même lignée, (Gesbert, 2005) et (Abadie, 2012) se sont intéressés à la description formelle des spécifications des données géographiques afin d'améliorer leur intégration. (Gesbert, 2005) propose une formalisation des spécifications de bases de données géographiques avec un modèle ad-hoc mais qui permet d'exprimer les différentes règles de sélection et de représentation des entités géographiques du monde réel à inclure dans une base de données géographiques. Dans le cadre de notre travail, nous notons principalement à la proposition de (Abadie, 2012) d'un modèle formel défini dans le langage standard du Web sémantique OWL2 pour représenter ces mêmes règles de sélection et de représentation. Nous notons principalement la proposition qui introduit le concept « *Feature* » de l'ontologie **Dolce**[83] pour la représentation de la modélisation géométrique. Ce concept est différent de celui de « *Feature* » utilisé dans le cadre des normes de la série ISO 19100. Il représente dans ce cas précis « une entité physique « parasite » dont l'existence dépend de celle de son « hôte » ». Pour les données géographiques (Abadie, 2012) spécialise ce concept en « SpatialFeature ». Pour permettre d'expliciter la modélisation géométrique, le modèle se base sur les travaux de (Schade, 2010) et (Smith et Mark, 1998) pour définir un ensemble « d'éléments caractéristiques de la forme » des entités géographiques du monde réel, tels que perçus dans un contexte cartographique.

Représentation de la modélisation géométrique dans le vocabulaire de la sémantique des XY

Dans le cadre de notre travail, nous reprenons le concept « d'élément caractéristique de la forme » ainsi que celui d'« hôte » sans lequel le premier ne peut être défini. Nous appliquons cela à un niveau fin, car dans notre modèle, chaque géométrie peut être décrite par sa modélisation géométrique (voir Figure 3.10). Nous formalisons les différents types d'éléments caractéristiques de la forme d'une entité géographique du monde réel (tels que définis par (Shade, 2010) et (Smith et Mark, 1998)) dans une taxonomie SKOS.

Les éléments caractéristiques de la forme des entités géographiques de type **Bona Fide** sont des éléments qui peuvent être identifiés grâce à une discontinuité physique du terrain, comme le contour d'un toit ou le bord d'une route par exemple. Au contraire, les éléments de type **Fiat** correspondent à des éléments définis par des choix subjectifs ou arbitraires, comme les limites administratives ou les parcelles cadastrales. Les éléments caractéristiques **implicites** sont des éléments qui peuvent être obtenus grâce à des opérations géométriques appliquées sur des éléments caractéristiques de la forme des entités géographiques de type **Bona Fide**, comme l'axe d'une route par exemple ou l'enveloppe convexe d'une surface. La Figure 3.11 montre un exemple de représentation de la modélisation géométrique d'une géométrie associée à une ressource géographique.

La géométrie peut être saisie par rapport à une entité du monde réel de type différent de celui l'objet géoréférencé. Par exemple, les adresses peuvent être géoréférencées par les centroïdes de bâtiments : l'hôte de l'élément caractéristique de la forme dans ce cas est de type « Bâtiment ». Il peut également s'agir d'une sous-partie de l'objet géoréférencé. Par exemple, les bâtiments des

[83] http://www.loa.istc.cnr.it/old/DOLCE.html

bases bases de données de l'IGN sont géoréférencés par les contours de leurs toîts: l'hôte de l'élément caractéristique de la forme dans ce cas est de type « Toît ».

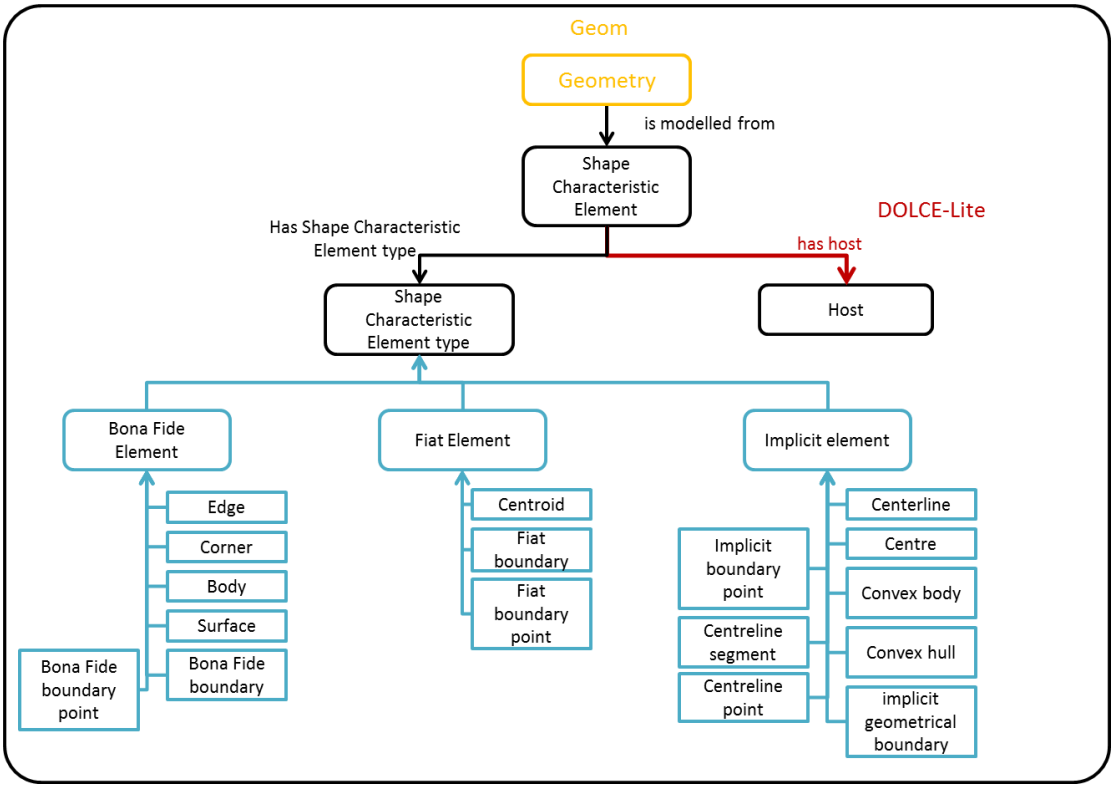


Figure 3.10 Extrait du vocabulaire de la sémantique des XY pour représenter la modélisation géométrique.

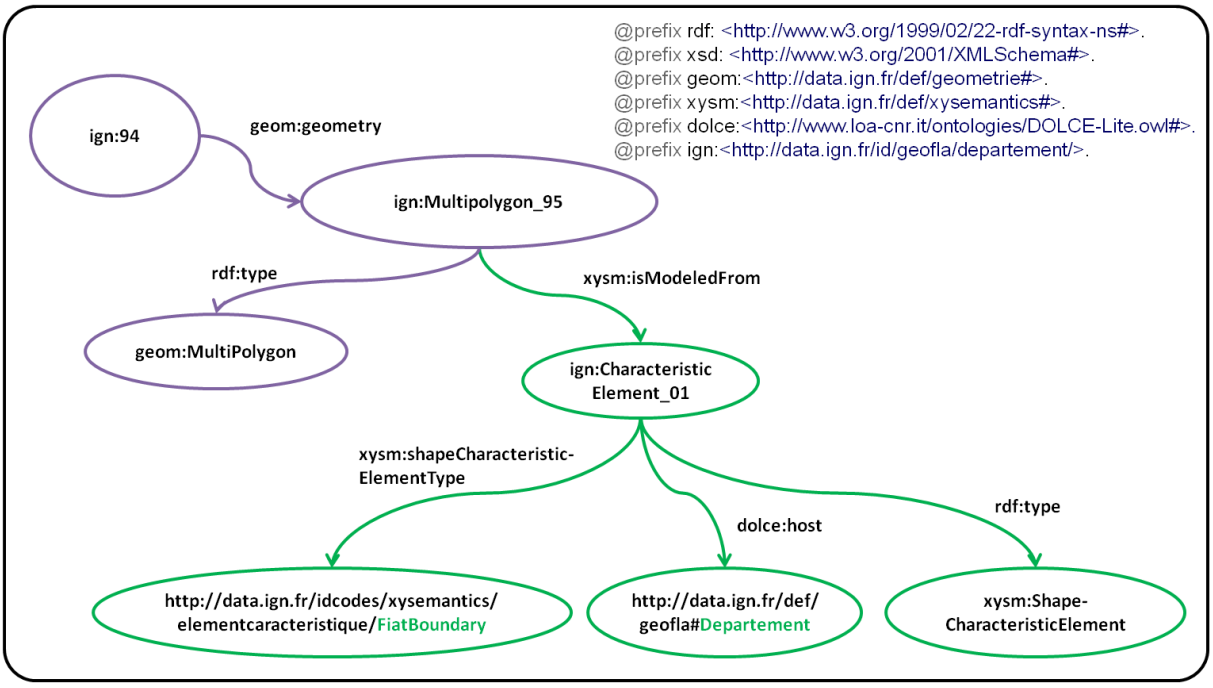


Figure 3.11 Représentation de la modélisation géométrique de la géométrie délimitant le département 94 dans la source Geofla http://data.ign.fr/id/geofla/departement/94

3.1.3 Métadonnée sur le caractère vague de certaines entités géographiques

Il s'agit de disposer de connaissances permettant de savoir si la géométrie concernées représente une entité géographique bien délimitée (comme un bâtiment par exemple) ou d'une entité géographique vague (comme les vallées ou les montagnes par exemple). Dans notre modèle, nous pouvons renseigner cette information au niveau de chaque géométrie en choisissant un élément caractéristique de la forme de type « élément vague » (voir Figure 3.12). Ainsi, cette information peut être prise en compte dans la comparaison géométrique.

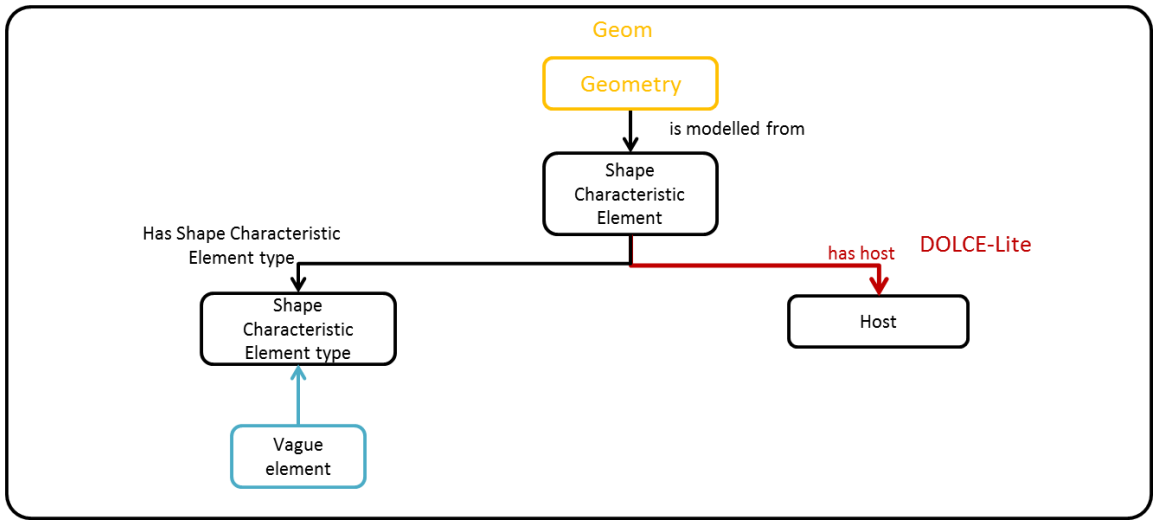


Figure 3.12 Extrait du vocabulaire de la sémantique des XY pour représenter le caractère vague des éléments caractéristiques de la forme de certains types d'entités géographiques.

3.1.4 Métadonnée sur la résolution géométrique

Dans les bases de données géographiques, la résolution géométrique donne l'ordre de grandeur géométrique des phénomènes représentés dans la base et fournit une compréhension de la densité des données. Comme présenté avant (*cf.* section 1.1.3), la notion de résolution est souvent liée aux données Raster, car elle correspond à la dimension représentée sur le terrain par un pixel. Bien que cette notion soit ambiguë pour les données vectorielles (Ruas, 2002), elle est souvent fournie pour faire référence à la résolution de la source utilisée pour la création des données, ou la résolution d'une carte comparable.

En effet, selon la norme ISO 19115-1 l'élément de métadonnée **MD_Resolution** est introduit pour spécifier la résolution géométrique des données. Selon cette norme, la résolution (sur un plan) doit être représentée comme: une échelle équivalente, une distance horizontale au sol ou une distance angulaire. L'échelle équivalente désigne l'échelle d'une carte ou d'un plan papier d'un niveau de détail comparable à celui des données. C'est le cas par exemple de la résolution fournie dans les métadonnées de livraison de la BD TOPO® (voir Figure 3.13). La distance horizontale au sol représente la taille d'échantillonnage qui constitue la plus petite longueur entre deux points lors de la saisie des géométries des objets géographiques. La distance angulaire correspond à la mesure d'échantillonnage angulaire qui constitue le plus petit angle entre deux segments lors de la saisie des géométries des objets géographiques.

Edition des données d'origine:	
– Edition	103
– Date	2010-09-22
Résolution spatiale :	
– Echelle équivalente	1:10000
Organisation responsable de la ressource :	
– Organisation	
▪ *Nom*	Institut Géographique National (IGN-F)
▪ *Mél*	sav.bd@ign.fr
– Rôle	pointOfContact

Figure 3.13 Résolution géométrique représentée par une échelle équivalente. Extrait des métadonnées de la BD TOPO®. (http://professionnels.ign.fr/sites/default/files/métadonnées de produit BD TOPO®.html)

Nous proposons donc dans notre vocabulaire d'adopter directement la proposition de la norme ISO 19115-1 en incluant un élément de résolution géométrique (voir Figure 3.14) qui se spécialise en trois classes : une classe de résolution par distance au sol décrite par sa valeur et son unité, une classe de résolution par distance angulaire décrite également par sa valeur et son unité, et une classe de résolution par échelle équivalente décrite par la valeur son dénominateur.

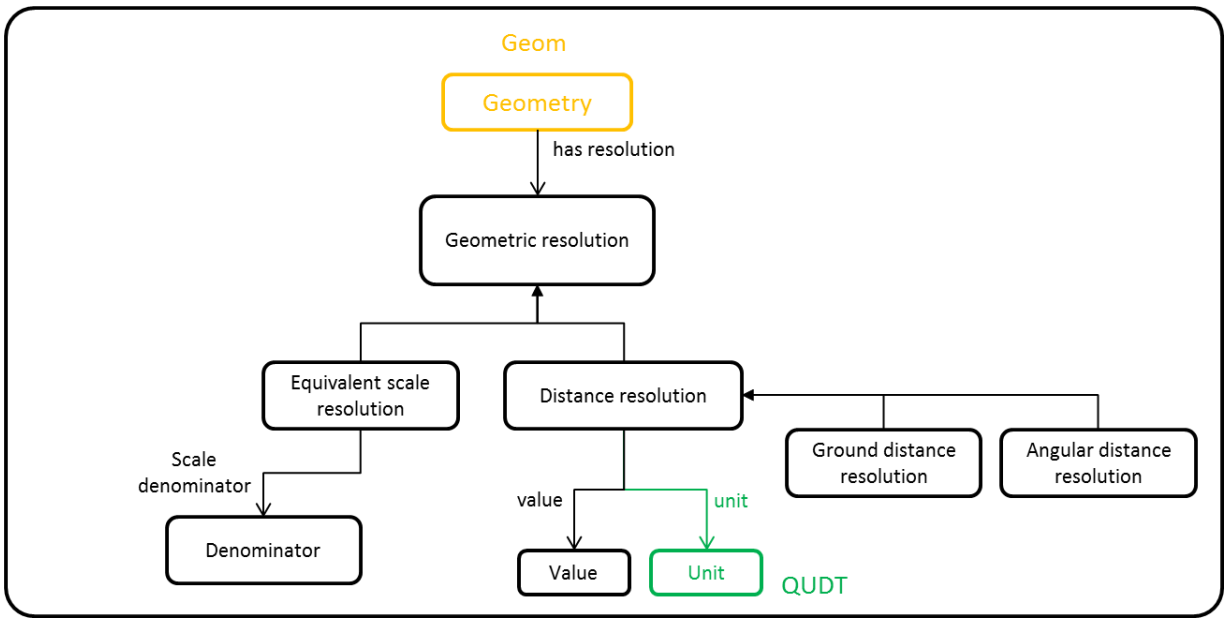


Figure 3.14 Extrait du vocabulaire de la sémantique des XY pour représenter la résolution géométrique.

Ainsi, la résolution géométrique d'un département de la base Geofla®, dont les métadonnées de livraison précisent qu'elle est d'une échelle équivalente de 1:100000, peut tout simplement être représentée comme montré dans la Figure 3.15.

Nous présentons dans la suite comment peupler le vocabulaire de la sémantique des XY au niveau de chaque géométrie d'un jeu de données.

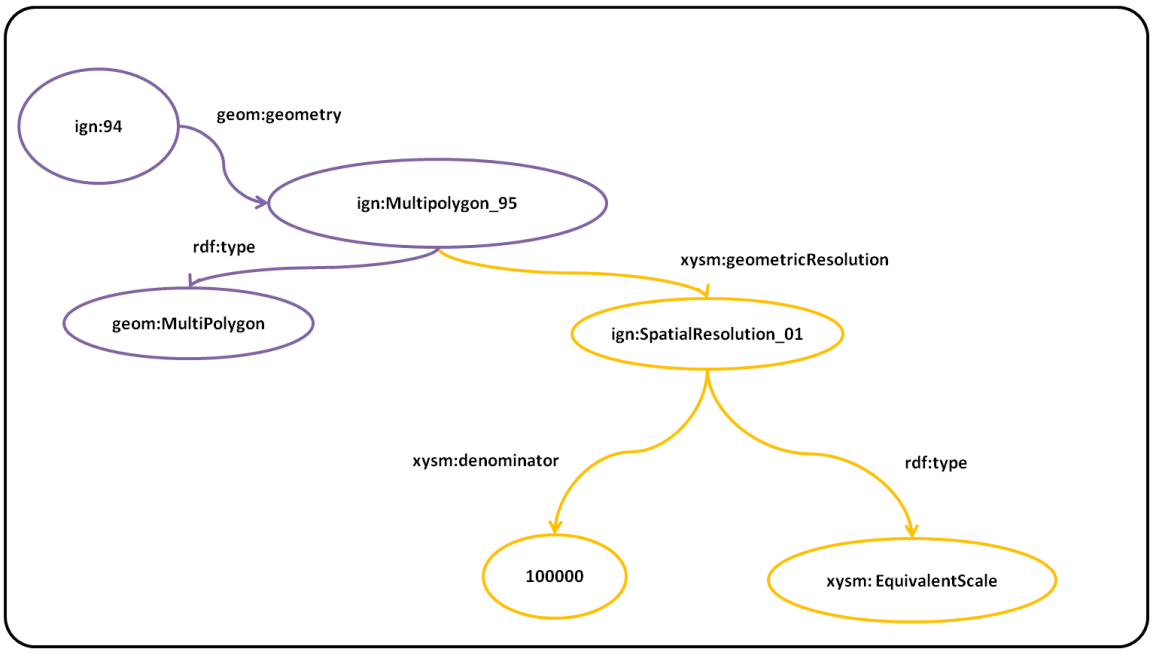


Figure 3.15 Exemple de représentation en RDF de la résolution géométrique de la géométrie délimitant le département 94 dans la source Geofla® http://data.ign.fr/id/geofla/departement/94

3.2 Peuplement du vocabulaire de la sémantique des XY

Le vocabulaire de la sémantique des XY est modélisé pour permettre la représentation de métadonnées sur des données géoréférencées au niveau le plus fin, à savoir leurs géométries. Instancier ces métadonnées pour un jeu de données est une tâche qui peut s'avérer plus compliquée que la tâche de représentation des métadonnées générales de tout un jeu de données. Les sources de données géographiques de référence sont souvent fournies avec des fiches descriptives et des métadonnées y compris des métadonnées sur les géométries. En outre, ces sources de données garantissent une homogénéité interne des caractéristiques des géométries : pour une même classe, les géométries ont le plus souvent les mêmes modélisations géométriques, la même précision planimétrique, etc. Même dans certains cas où les géométries d'une même classe ont des caractéristiques différentes (*ex.* une précision planimétrique qui varie selon les instances), des indications supplémentaires sont fournies pour préciser les différents cas de figure (*ex.* les différentes précisions planimétriques possibles des adresses de la BD ADRESSE® sont précisé les différents types de localisation *cf.* Figure 3.4). En revanche, les métadonnées ne sont fournies que rarement avec les jeux de données du Web. En effet, le Web de données regorge de sources de données géoréférencées qui ne fournissent aucune indication sur les méthodes utilisées pour le géoréférencement, ni sur sa qualité. De plus, les sources de données d'origine collaborative ne présentent pas forcément des consignes sur la manière de représenter les références spatiales, et même si elles le font, elles ne garantissent pas forcément le respect de ces consignes par les contributeurs. Nous décrivons dans la suite de cette partie comment peupler le vocabulaire dans les deux cas de figure, *c.-à-d.* en présence en absence de métadonnées sur les caractéristiques des géométries.

3.2.1 Peuplement du vocabulaire dans le cas où les métadonnées sur les géométries sont fournies avec les données

Les métadonnées descriptives des données géographiques sont souvent fournies sous forme de fichiers dans des formats textuels ou XML. Par exemple, ceux fournis avec les données de référence de l'IGN, facilitent amplement la tâche de peuplement du vocabulaire. Il suffit dans ce cas de traduire ces métadonnées selon le modèle RDF en utilisant notre vocabulaire et de les associer aux géométries adéquates. Ceci peut être facilement réalisable grâce à des requêtes Sparql d'insertion de données.

Prenons l'exemple des données BD ADRESSE®2.2 qui est l'une des composantes du Référentiel à Grande Échelle[84] (RGE) de l'IGN. Dans la classe ADRESSE de cette base, les adresses sont géoréférencées par des points. Selon les fichiers descriptifs du contenu de cette base, les points adresses sont localisés selon différentes modélisations géométriques possibles selon les cas: à la plaque-adresse, à l'entrée du bâtiment, à 4.5m de l'axe de la voie (par projection à partir des centroïdes des parcelles, par interpolation ou arbitrairement), sur une zone d'adressage ou encore au centre de la commune. Le type de modélisation géométrique est renseigné pour chaque adresse grâce à un attribut **TYPE_LOC**. Le descriptif de contenu présente également les différentes précisions planimétriques des géométries des différentes classes de la base selon la source de données utilisée (*cf.* Figure 3.2). Contrairement aux autres classes de la BD ADRESSE® (ROUTE_ADRESSE, CHEF_LIEU, LIEU_DIT_HABITE, etc.) la classe ADRESSE n'a pas d'attribut **PREC_PLANI,** pour renseigner la précision planimétrique de chaque objet. Cependant, le document associe à chaque type de localisation des mesures obtenues par des contrôles de qualité effectués sur un échantillon qui comprend deux départements (45, 76) et quatre agglomérations (Grand Lyon, Communauté Urbaine de Toulouse, Rennes et Poitiers). Les résultats de cette évaluation sont détaillés dans la Figure 3.16.

Zone			Valeurs (m)
Toutes les données	Distance	Moyenne	25
		Ecart type	150
Type de localisation	Plaques	Moyenne	8
		Ecart type	12
	Interpolation Tronçon	Moyenne	40
		Ecart type	90
	Interpolation voie	Moyenne	90
		Ecart type	230
	Projetée du centre parcelles	Moyenne	15
		Ecart type	25
	A la zone d'adressage	Moyenne	300
		Ecart type	400

Figure 3.16 Extrait du descriptif de contenu de la BD ADRESSE®2.2 pour les résultats du contrôle de qualité géométrique.

Le descriptif du contenu ne fournit pas d'autres détails sur la méthode utilisée pour l'évaluation de cette précision ni sur les sources de données externes utilisées dans ce cadre. Le descriptif de

[84] http://professionnels.ign.fr/rge

contenu de la BD ADRESSE® ne fournit pas directement d'information sur la résolution des données. Cependant, une fiche de métadonnées générales fournie avec les livraisons de la BD ADRESSE® spécifie que les données sont pour une utilisation optimale à une résolution d'une échelle équivalente de 1 : 10000. Tout ceci permet donc d'associer à la géométrie de chaque instance de la classe ADRESSE une valeur de précision planimétrique (établie d'après la valeur de l'attribut **TYPE_LOC**) ainsi qu'une valeur de résolution.

Nous avons donc converti les données de la BD ADRESSE® de Paris ainsi que leurs métadonnées en RDF. Les données fournies par l'IGN peuvent être récupérées sous le format de fichier Shapefile[85]. Nous avons transformé es données d'adresses de ce format vers le modèle RDF. Pour cela nous avons utilisé la plateforme Datalift[86]. Cette plateforme permet de réaliser les différentes étapes d'un processus de publication de données selon les bonnes pratiques du Web de données. Elle permet notamment la transformation de données à partir d'une variété de formats (bases de données relationelles, CSV, XML, Shapefile, etc.) en données RDF. La plateforme permet aussi de définir des URIs de base pour l'identification des différentes ressources ainsi que de choisir des vocabulaires pour typer et structurer les données. En ce qui concerne les données géographiques vectorielles, les géométries sont représentées conformément au vocabulaire Geom[87]. Ainsi, nous avons pu transformer les fichiers Shapefile des adresses de Paris en données RDF dotées de ressources de type « geom:Geometrie ». Associer les métadonnées à chaque géométrie est effectué grâce à plusieurs requêtes Sparql d'insertion. Chaque requête a pour but d'insérer les métadonnées des géométries pour un type de modélisation géométrique bien précis identifiable dans les données d'origines grâce aux valeurs de l'attribut TYPE_LOC.

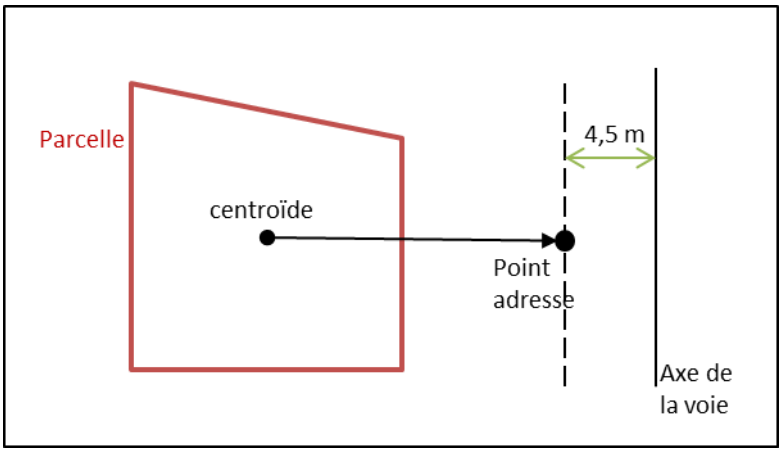


Figure 3.17 Exemple de positionnement des points adresse dans la BD ADRESSE® : par projection à partir du centroïde de la parcelle

Prenons par exemple les adresses dont le type de localisation est « Projection », ce qui signifie que l'adresse est localisée par un point obtenu par projection du centroïde de la parcelle sur le bord de la voie correspondante que l'on assimile à une polyligne imaginaire située à 4.5m de l'axe de la voie, comme présenté dans la Figure 3.17.Nous interprétons la modélisation géométrique dans ce cas en

[85] Format de fichier de données propriétaire de l'entreprise ESRI. Il permet une représentation vectorielle des données géographiques.
[86] http://www.datalift.org/
[87] http://data.ign.fr/def/geometrie

associant à chaque géométrie un élément caractéristique de la forme de type « *Implicit Boundary Point* » (point d'une limite implicite), associé à un élément hôte de type «Route» (*ex.* la classe topo[88]:Route de l'ontologie de la BD TOPO®). La précision planimétrique des géométries est décrite par une méthode d'évaluation de type « *Sample Based Evaluation* » (évaluation par usage d'échantillon) et un résultat d'évaluation de valeur 15 et d'unité « mètre » (Qudt:meter). Ce résultat correspond à la valeur moyenne des précisions planimétriques des géométries obtenues par projection à partir des centroïdes des parcelles (*cf.* Figure 3.16). Conformément aux métadonnées générales de la BD ADRESSE®, la résolution des géométries est de type « *Equivalent Scale Resolution* » et est donc décrite par un dénominateur de valeur 10000. L'annexe A détaille les différentes requêtes effectuées pour associer les métadonnées sur les caractéristiques des géométries à leur géométrie associée en fonction de la valeur de l'attribut TYPE_LOC des instances de la classe ADRESSE.

3.2.2 Peuplement du vocabulaire en l'absence de métadonnées sur les géométries fournies avec les données

Nous avons vu dans la partie 1.1.3 que les sources de données du Web peuvent rassembler des ressources géoréférencées d'origines diverses (données géographiques de producteurs traditionnels, données issues de collectes participatives, etc.). Dans le cas des données géoréférencées issues de données de référence fournies par des producteurs traditionnels, les métadonnées sont souvent livrées avec les données comme nous l'avons vu dans la partie précédente. Cependant, pour les données provenant d'autres origines, les références spatiales ne sont pas forcément saisies pour un niveau de détail fixe et selon des spécifications uniques et bien définies, surtout dans le cas de données issues de données collaboratives. De plus, même quand les conditions de saisie des géométries sont connues, elles ne sont pas forcément Dans ce cas, peupler le vocabulaire de la sémantique des XY devient une tâche compliquée. Les caractéristiques des géométries, notamment la modélisation géométrique et la précision planimétrique, peuvent varier significativement d'une ressource à une autre. Leur évaluation constitue une tâche laborieuse si elle est réalisée manuellement pour chaque géométrie dans une source de données. Nous proposons donc d'automatiser cette tâche.

Approches existantes pour l'évaluation des caractéristiques des géométries

L'évaluation de la précision planimétrique des géométries rentre dans le cadre général des travaux d'évaluation de la qualité des données géographiques. Les bases de données géographiques des fournisseurs traditionnels de données sont produites par un processus de saisie conçu pour garantir un niveau de détail défini en amont. Selon la norme ISO 19157, une évaluation de la qualité doit ensuite être appliquée pour vérifier la conformité des géométries saisies au niveau de détail défini et aux spécifications de saisie, ainsi que pour renseigner la qualité effective des données. La précision planimétrique, comme tout autre élément de la qualité des données géographiques, peut être évaluée selon la norme ISO 19157 d'une manière indirecte, par déduction à partir de connaissances sur les données comme leur généalogie par exemple, ou d'une manière directe par une inspection complète des données ou d'un échantillon représentatif de celles-ci (*cf.* section 3.1.1). Les méthodes

[88] http://data.ign.fr/def/topo#

d'évaluation directes consistent à inspecter les éléments du jeu de données par analyse de la « proximité des valeurs des coordonnées des géométries par rapport aux valeurs vraies ou reconnues en tant que telles» (méthode absolue), ou par analyse de la « proximité des positions relatives des objets par rapport à leurs positions relatives respectives vraies ou reconnues en tant que telles» (méthode relative). Les méthodes absolues nécessitent de disposer d'un jeu de données de référence de qualité supérieure et maîtrisée pour l'élément de qualité à évaluer. Il est nécessaire donc de connaître en amont les modélisations géométriques des géométries à évaluer pour pouvoir calculer leurs écarts par rapport à leurs positions homologues dans le jeu de données de référence. À notre connaissance, il n'existe pas d'approches qui permettent d'identifier automatiquement la modélisation géométrique des géométries associées aux ressources publiées sur le Web.

Le catalogue des mesures standardisées de la norme ISO 19157 propose de nombreuses méthodes numériques d'évaluation de la précision planimétrique des géométries. Toutefois, ces méthodes ne sont applicables qu'à des jeux de données de précision planimétrique supposée homogène. Dans le cas contraire, la norme conseille de catégoriser les données par cause d'hétérogénéité et d'appliquer indépendamment la méthode choisie à chaque sous-ensemble. L'évaluation de la qualité des informations de localisation produites par des projets participatifs fait l'objet de nombreux travaux. (Hauff, 2013) propose une approche pour évaluer la précision planimétrique des geotags associés aux images FlickR concernant des bâtiments remarquables. Celle-ci repose sur une annotation préalable des images de l'échantillon évalué : photographie prise à l'intérieur du bâtiment, à l'extérieur ou dans un lieu de prise de vue inconnu. Le choix de cette typologie est guidé par l'intuition selon laquelle les photographies prises à l'intérieur d'un bâtiment devraient être localisées plus précisément que les autres. L'estimation de la précision planimétrique des geotags est réalisée par un calcul des distances moyennes entre les geotags FlickR des photographies de chacune de ces trois classes et les coordonnées fournies par l'article Wikipedia décrivant le bâtiment photographié. Le choix de Wikipedia comme référence pour les localisations des bâtiments n'est cependant pas discuté. OpenStreetMap est probablement la source de données géographiques participatives dont la qualité a fait l'objet du plus grand nombre de travaux. (Girres et Touya, 2010) évalue la qualité des données OpenStreetMap sur le territoire français à l'aide d'un ensemble de mesures directes standard. L'estimation de la précision planimétrique des données est réalisée en calculant les décalages moyens entre trois échantillons de données OpenStreetMap (respectivement dotés de géométries ponctuelles, linéaires et surfaciques) et leurs homologues extraites du Référentiel à Grande Échelle. (Barron et al., 2014) propose un ensemble de mesures d'évaluation indirectes de la qualité des données OpenStreetMap, fondées sur les métadonnées de généalogie disponibles. Les travaux sur l'évaluation de la qualité des données liées publiées sur le Web portent principalement sur leur conformité aux bonnes pratiques du Web de données. Ainsi, l'état de l'art sur les mesures d'évaluation de la qualité des données du Web proposé par (Zaveri et al., 2015) présente en priorité les mesures sur les questions de déréférencement, de licence et d'interconnexion. Les mesures concernant la qualité intrinsèque des données, leur adéquation à un besoin et leur représentation viennent ensuite. Or, aucune des mesures présentées ne s'at tâche spécifiquement aux caractéristiques spatiales des ressources. Toutefois, l'approche proposée par (Wienand et Paulheim, 2014) pour la détection de valeurs numériques aberrantes dans DBpedia permet l'identification de valeurs de coordonnées ou d'altitudes invraisemblables. De plus, (Ahlers, 2013) évalue la précision planimétrique des coordonnées associées aux ressources du gazetier GeoNames en termes de nombre de décimales transcrites dans ces coordonnées.

Nous proposons ici une approche en deux étapes pour l'extraction des caractéristiques des géométries. Nous proposons tout d'abord une méthode pour l'identification automatique des différentes modélisations géométriques des références spatiales des ressources. Il s'agit de retrouver, pour chaque ressource à l'intérieur d'une même source de données, quel choix d'élément caractéristique de sa forme a été fait lors de la saisie des coordonnées utilisées pour la localiser. Identifier l'élément caractéristique de la forme pour chaque géométrie permet ensuite d'évaluer la précision planimétrique des références spatiales des ressources en adaptant les méthodes d'évaluation directes de la précision planimétrique absolue des données géographiques, et en s'inspirant des différentes méthodes d'évaluation de la précision appliquées aux données participatives.

Une méthode d'identification des modélisations géométriques des géométries basée sur des données géographiques de référence

Les choix de modélisation géométrique qui ont été faits lors de la saisie des géométries des données du Web sont à notre connaissance très rarement renseignés avec les données. L'ajout de ressources géoréférencées à certains jeux de données, notamment dans le cadre de plateformes de saisie participatives, peut être libre de toute restriction sur ce choix. C'est le cas par exemple de l'ajout manuel de ressources à la source GeoNames, dont le manuel d'utilisation[89] ne donne pas de consigne sur la manière précise de choisir la localisation d'une ressource par rapport à la forme de l'entité du monde réel correspondante perçue sur le fond cartographique ou orthophotographique utilisé pour la saisie. Même pour les sources de données où des recommandations de saisie des géométries ont été fournies, on ne peut directement savoir si ces recommandations ont été respectées ou pas. C'est le cas par exemple des ressources de DBpedia qui sont majoritairement issues de Wikipedia : des recommandations de saisie des coordonnées ont été proposées dans le cadre du Projet «*WikiProject Geographical Coordinates*»[90], mais ne sont pas forcément respectées. La Figure 3.18 montre un exemple de localisations de ressources qui décrivent des monuments parisiens sur DBpedia, où la recommandation est de saisir les coordonnées à l'entrée des sites concernés. On constate rapidement que nombre de monuments sont localisés au centre du bâtiment correspondant.

Ainsi, que des recommandations de saisie soient formulées ou pas, une analyse des géométries des ressources s'impose afin d'identifier les différentes modélisations géométriques effectivement choisies lors de la saisie des données. Nous proposons une approche, de type « rétro-ingénierie », qui part des données pour identifier les différentes modélisations géométriques des références spatiales. Nous partons de l'hypothèse principale qu'une géométrie d'une ressource découle d'un choix intentionnel de modélisation géométrique par le contributeur[91] qui l'a saisie. Nous proposons donc de formuler différentes hypothèses sur les choix faits par les contributeurs lors de la saisie des géométries des ressources. Ces choix hypothétiques peuvent être déterminés par analyse empirique visuelle des références spatiales relativement à des données ou des fonds cartographiques de référence. Plus précisément, cette analyse relative consiste à comparer visuellement les références

[89] http://www.geonames.org/manual.html
[90] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geographical_coordinates
[91] Nous utilisons ici le terme « contributeur » pour désigner la personne ayant saisi la géométrie associée à une ressource, que celle-ci soit issue d'une collecte de données participative ou d'un processus de saisie institutionnel classique

spatiales des ressources aux géométries utilisées pour représenter des entités géographiques de types sémantiquement proches ou équivalents au sein d'un jeu de données géographiques de référence. Dans la Figure 3.18 par exemple, comme il s'agit de monuments, on peut éventuellement comparer leur localisation avec celles des bâtiments et autres constructions visibles sur une orthophotographie haute-résolution. L'analyse visuelle permet ainsi d'identifier les différentes tendances des choix hypothétiques de modélisation géométrique qui émergent des données. Il demeure cependant difficile, voire impossible, de vérifier que ces choix probables correspondent aux choix intentionnels des contributeurs.

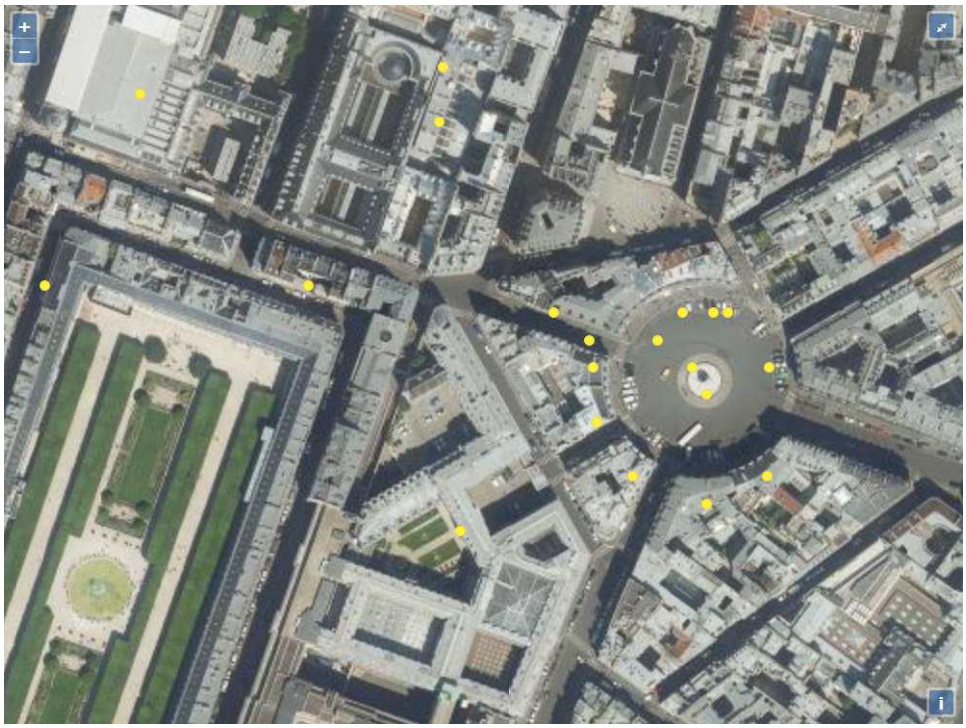


Figure 3.18 Co-visualisation des localisations des monuments historiques de DBpedia (en jaune) avec un fond orthophotographique IGN

L'étape suivante consiste à associer chaque référence spatiale à l'un des types de modélisation géométrique identifiés. Réaliser cette tâche manuellement est très coûteux dans un contexte de données d'origines diverses et volumineuses, comme c'est le cas avec les sources de données du Web. Automatiser cette tâche, qui consiste à affecter des géométries à des types (**des classes**) de modélisations géométriques déjà identifiées, peut être réalisé grâce à une **classification par apprentissage supervisé**.

De façon générale, l'apprentissage consiste à construire des modèles fiables et précis de prédiction à partir de connaissances ou d'expériences déjà acquises (Mohri et al., 2012). La classification par apprentissage a pour objectif d'identifier les classes auxquelles appartiennent des objets à partir de leurs caractéristiques ou attributs descriptifs. Plus précisément, on parle de classification par apprentissage supervisé lorsqu'on part d'un ensemble de classes déjà définies et qu'on dispose d'exemples de chaque classe. Ceci correspond à notre contexte ou les différentes classes sont prédéfinies par analyse visuelle des données. Une méthode de classification par apprentissage supervisé peut être définie comme suit (Campedel , 2005): soit P la population d'objets à classifier et

soit D l'ensemble de leurs descripteurs. Soit $\{C_1, \dots, C_k\}$ l'ensemble des classes prédéfinies. Soit la fonction $X: P \rightarrow D$ qui associe une description à tout élément de P. Soit la fonction $Y: P \rightarrow \{C_1, \dots, C_k\}$ qui associe une classe à tout élément de P. Cette fonction représente l'association de chaque objet de la population à sa classe effective. L'objectif est d'apprendre $C: D \rightarrow \{C_1, \dots, C_k\}$ la fonction dite de classement ou procédure de classification, de façon à ce que $C(X)$approche au mieux Y. La fonction C doit permettre donc de classifier les objets selon les valeurs de leurs descripteurs. Apprendre cette fonction d'une manière supervisée revient à dire qu'on part d'un ensemble de données d'apprentissage $S \times \{C_1, \dots, C_k\}$ obtenu grâce à la fonction $A: S \rightarrow \{C_1, \dots, C_k\}$ telles que $S \subset P$ et $A(S) = Y(S)$. Ceci revient à dire que l'ensemble d'apprentissage rassemble des exemples d'objets avec leurs classes effectives. L'ensemble d'apprentissage doit impérativement contenir des exemples de toutes les classes $\{C_1, \dots, C_k\}$. Les deux défis principaux auxquels une méthode d'apprentissage doit faire face sont le choix d'un ensemble de descripteurs D qui soient pertinents et le choix d'un ensemble d'apprentissage $S \times \{C_1, \dots, C_k\}$ qui soit le plus représentatif possible des différentes classes (Mohri et al, 2012). Une multitude d'approches ont été proposées dans l'état de l'art afin de permettre la définition de la fonction de classification à partir d'un ensemble d'apprentissage. Ce travail n'a pas pour but de concevoir une nouvelle approche d'apprentissage, mais de définir un cadre dans lequel les algorithmes d'apprentissages existants peuvent être appliqués pour l'identification automatique des modélisations géométriques des références spatiales associées aux ressources sur le Web de données.

Ainsi, nous formalisons notre problème comme suit : nous disposons d'une population de géométries G et d'un ensemble de classes de types de modélisations géométriques $\{C_1, \dots, C_k\}$. Nous devons définir un ensemble de descripteurs pertinents D et sélectionner un ensemble d'apprentissages $S \times \{C_1, \dots, C_k\}$ (avec $S \subset G$) afin de pouvoir définir la fonction de classification C. Pour que les descripteurs D soient pertinents, ils doivent constituer des indicateurs descriptifs dont la combinaison des valeurs permet de discriminer les différentes classes. Pour les définir, nous proposons de partir de la même intuition qui nous a permis d'identifier les classes en premier lieu, c'est-à-dire de procéder par analyse des références spatiales des ressources relativement aux géométries qui représentent des entités géographiques de types sémantiquement proches ou équivalents au sein d'un jeu de données géographiques de référence. Les descripteurs doivent donc permettre d'expliciter la position relative des références spatiales par rapport aux données géographiques de référence. Il peut s'agir d'une distance ou d'une relation entre les références spatiales et les éléments caractéristiques de la forme des entités géographiques représentées par les objets géographiques de référence. Par exemple, il peut s'agir de la distance entre les références spatiales et les axes des routes ou de l'inclusion des références spatiales dans les parcelles cadastrales. La sélection d'un ensemble d'apprentissage consiste à trouver pour chaque classe de modélisation géométrique un échantillon représentatif de références spatiales facilement identifiables pour cette classe. Une fois l'ensemble d'apprentissage défini, nous appliquons un algorithme d'apprentissage, pris parmi les algorithmes de l'tat de l'art, pour apprendre un modèle de classification qui permet d'affecter chaque référence spatiale à une classe de modélisation géométrique.

Évaluation de la précision planimétrique des géométries des ressources

Évaluer la précision planimétrique des géométries des ressources revient à calculer leurs écarts par rapport aux éléments caractéristiques de la forme identifiés comme correspondant aux choix

hypothétiques de saisie des contributeurs lors de l'étape précédente et repérés au sein d'un jeu de données géographiques de référence de précision planimétrique absolue supposée meilleure et connue. La valeur de précision planimétrique affectée à chaque géométrie est estimée par la somme de son écart par rapport à son élément caractéristique de la forme correspondant avec la précision planimétrique de la géométrie de l'objet hôte de cet élément caractéristique de la forme. Par exemple si une ressource est géoréférencée par un point qui a été classifié comme étant saisi à l'axe d'une route, la précision planimétrique de ce point est estimée comme la somme de la distance de ce point à l'axe de la route correspondante et de la précision planimétrique de cette route.

Mise en œuvre de l'approche d'identification automatique de la modélisation géométrique et d'évaluation de la précision planimétrique des géométries d'un jeu de données

La description de la mise en œuvre dans cette partie portera sur deux exemples choisis afin de démontrer deux points essentiels. Le premier est la faisabilité de l'approche d'identification des modélisations géométriques en utilisant des données géographiques de référence. Le deuxième est l'utilisation des résultats de cette première approche pour l'estimation de la précision planimétrique des références spatiales. Le premier exemple concerne le jeu d'adresses du Grand Lyon alors que le deuxième concerne les ressources du chapitre français de DBpedia qui décrivent les monuments historiques de Paris.

Cas du jeu de données d'adresses du Grand Lyon

Afin d'évaluer la faisabilité de l'approche d'identification automatique de la modélisation géométrique, nous avons décidé de l'appliquer d'abord sur un jeu de données contenant plusieurs modélisations géométriques, mais qui sont connues en amont, pour vérifier si l'approche nous permet de classer correctement chaque référence spatiale dans sa classe de modélisation géométrique effective. Pour cela, nous avons utilisé les données des adresses du Grand Lyon (*cf.* section 3.1.2). Dans ce jeu de données chaque adresse est fournie avec trois modélisations géométriques différentes : point implanté sur le bâtiment (vers son barycentre), point implanté sur l'axe de la voie correspondante et point implanté à la position exacte du numéro (*c.-à-d.* de la plaque) d'adresse (*cf.* Figure 3.9). Nous avons extrait un jeu de données à partir de cette base d'adresses en sélectionnant aléatoirement une seule modélisation géométrique pour chaque adresse (voir Figure 3.19).

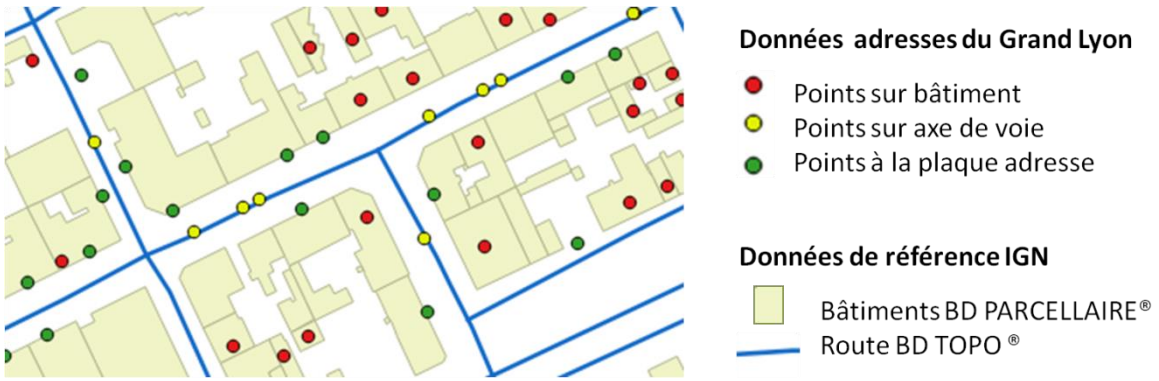


Figure 3.19 Extrait des géométries des adresses du grand Lyon à co-visualisées avec des données géographiques de référence de l'IGN

Dans cet exemple, nous avons sélectionné les 7014 géométries des adresses appartenant au troisième arrondissement de la ville de Lyon (code Insee 69383). Ces géométries constituent donc la population G à classifier. Les classes prédéfinies $\{C_1, C_2, C_3\}$ représentent respectivement les trois modélisations géométriques possibles : sur le bâtiment, sur l'axe de voie et à la plaque-adresse. Il nous est donc nécessaire de définir l'ensemble D des descripteurs d'apprentissage pertinents. Nous utilisons pour cela des jeux de données géographiques de référence de l'IGN, choisis pour mettre en évidence les différentes classes de modélisation géométriques (voir Figure 3.19): les bâtiments de la BD PARCELLAIRE® et les routes de la BD TOPO®. Nous avons remarqué que les points implantés sur les bâtiments ont tendance à se rapprocher des barycentres de ces derniers alors que les points qui sont implantés aux plaques adresse sont logiquement proches de la façade. Nous avons donc identifié quatre descripteurs que nous avons jugés pertinents par rapport aux différentes classes de modélisation géométrique (voir Figure 3.20): la distance entre le point adresse et la route (d_r), la distance entre le point adresse et la façade du bâtiment le plus proche (d_f), la distance entre le point adresse et le barycentre du bâtiment le plus proche (d_c) et l'inclusion ou non du point adresse dans la surface du bâtiment le plus proche (inc).

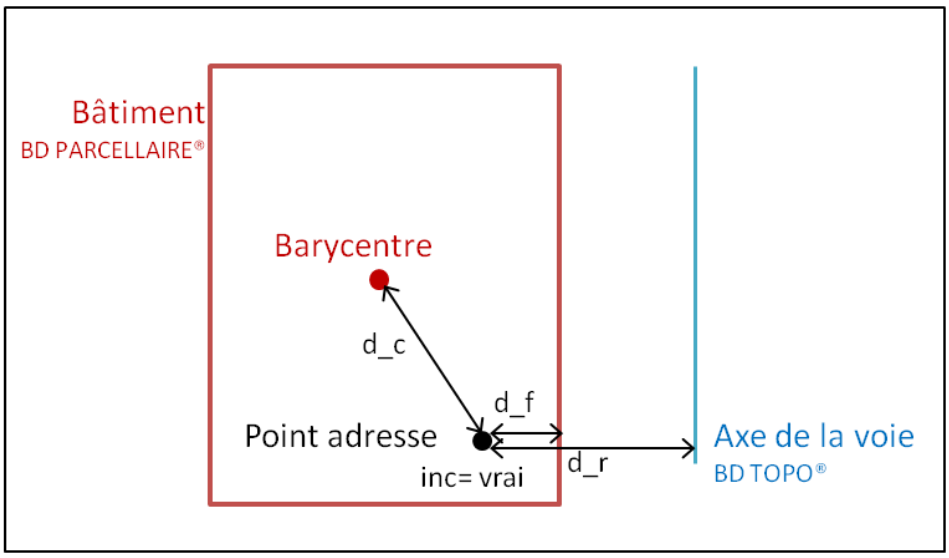


Figure 3.20 Descripteurs choisis pour l'apprentissage du modèle de classification des modélisations géométriques des adresses du Grand Lyon

L'étape suivante consiste à définir la fonction $X: G \rightarrow D$ qui associe une description à tout élément de G. Autrement dit, on doit calculer les valeurs des indicateurs d_r, d_f, d_c, inc pour chaque point adresse. Cette tâche peut être réalisée grâce aux outils classiques d'analyse spatiale. Dans notre cas, nous avons utilisé l'outil QGIS[92].

Les classes de modélisation géométrique des références spatiales étant connues en amont, la fonction $Y: G \rightarrow \{C_1, C_2, C_3\}$ qui associe chaque géométrie à sa classe effective est considérée donc bien définie. Pour apprendre le modèle de classification de la fonction $C: D \rightarrow \{C_1, C_2, C_3\}$, nous proposons de comparer plusieurs algorithmes d'apprentissage de l'état de l'art mis en œuvre grâce à l'outil Weka[93]. Weka est un logiciel qui fournit des implémentations pour les algorithmes

[92] Quantum GIS, outil de SIG libre et ouvert. http://www.qgis.org/fr/site/
[93] http://www.cs.waikato.ac.nz/ml/weka/

d'apprentissage supervisé les plus répandus. Il permet de sélectionner un jeu de test d'apprentissage indépendant de la population à classifier, ou d'en créer aléatoirement à partir d'une partie de cette population. Weka permet également d'évaluer le résultat d'apprentissage en comparant les résultats de classification par le modèle appris par rapport aux classes effectives de la population. Nous avons sélectionné un ensemble d'apprentissage contenant en total 200 adresses (entre 55 et 80 adresses de chaque classe). Le tableau 3.21 détaille les performances des algorithmes utilisés (voir (George-Nektarios, 2013) pour des détails sur les algorithmes utilisés dans ces tests).

Méthode	Précision	Rappel	F-measure
Bayes Network	94,6%	94,5%	94,5%
Jrip	98,0%	98,0%	98,0%
Decision Table	97,6%	97,5%	97,5%
Random Forest	100%	100%	100%

Tableau 3.1 Évaluation des résultats de la classification par apprentissage supervisé des géométries des adresses du Grand Lyon.

Discussion des résultats

Les résultats de classification des géométries dans l'exemple des adresses du Grand Lyon sont plutôt satisfaisants. Ce jeu de données a été construit par sélection aléatoire de données dans 3 versions des adresses du Grand Lyon, chacune présentant un choix de modélisation géométrique particulier. Il nous permet de tester nos hypothèses et valider notre proposition d'approche pour l'extraction de connaissances sur la modélisation géométrique des données par classification supervisée. Le jeu d'apprentissage a été également sélectionné d'une manière aléatoire, et ne représente que 0.03% de la population. Dans cet exemple, on arrive même à reproduire parfaitement la classification effective avec la méthode « Random Forest ». Ces résultats tendent à démontrer l'efficacité de notre approche.

Cas des monuments historiques de Paris sur DBpedia

Dans cet exemple nous partons d'un jeu de données dotés de géométries dont nous ignorons la modélisation géométrique et la précision planimétrique des géométries. Nous avons extrait un ensemble de 625 ressources décrivant les monuments historiques de Paris, géoréférencées par des points (coordonnées de longitude et latitude), à partir du chapitre français de DBpedia[94]. Plus précisément, nous avons sélectionné les ressources décrites par une propriété « dcterms:subject » de valeur « dbpedia-fr: Catégorie:Monument_historique_de_Paris », ainsi que celles liées à la catégorie « dbpedia-fr: Catégorie:Monument_parisien » via le chemin « skos: broader| dcterms: subject ». Comme expliqué précédemment, les recommandations du projet « *WikiProject Geographical Coordinates* » suggèrent que ce genre de ressources doivent préférablement être saisies vers l'entrée du bâtiment correspondant. Or, comme on a pu le constater sur la Figure 3.18, cette recommandation n'est pas forcément respectée. Nous proposons donc de formuler des hypothèses sur les choix faits par les contributeurs lors de la saisie des coordonnées des monuments. Tout comme dans l'exemple précédent des adresses du Grand Lyon, nous comparons ensuite ces coordonnées des ressources DBpedia aux géométries correspondant aux divers choix de modélisation géométrique possibles au sein d'un jeu de données géographiques de référence.

[94] http://fr.dbpedia.org/. Données extraites en décembre 2013

En affichant les points qui localisent les monuments historiques DBpedia sur un fond orthophotographique IGN, on distingue trois types de choix de localisation : à proximité du centre du bâtiment classé comme monument historique, à proximité de sa façade et enfin à proximité de l'axe de la rue devant sa façade. Les deux premiers choix de représentation correspondent respectivement à deux types de recommandations de saisie du projet « *WikiProject Geographical Coordinates* » : pour les entités géographiques d'emprise spatiale importante, localiser leur centre, pour les bâtiments, localiser leur entrée principale. Le troisième correspond à une pratique courante pour la saisie des bases de données d'adresses. Tout comme dans l'exemple précédent, nous disposons de trois classes de modélisation géométrique. Pour définir les descripteurs d'apprentissage, nous nous sommes appuyés sur les mêmes données de références que pour l'exemple précédent, à savoir les bâtiments de la BD PARCELLAIRE® et les routes de la BD TOPO®. Des tests initiaux avec les mêmes descripteurs que dans l'exemple précédent nous ont semblé insatisfaisants, surtout pour distinguer une modélisation géométrique par un point à la façade d'une modélisation par un point au barycentre si le bâtiment correspondant est petit par exemple. Nous avons remarqué que les tailles des bâtiments à Paris sont plus hétérogènes que celles des bâtiments de notre exemple sur Lyon. Nous avons donc pris la taille des bâtiments et leur distance par rapport aux axes routiers en compte en choisissant les quatre descripteurs illustrés dans la Figure 3.21 : le rapport **d_r/a**, le rapport **d_f/b** , le rapport **(2*d_c)/c** ainsi que l'attribut **inc** qui précise si le point du monument est situé à l'intérieur de la surface du bâtiment ou non.

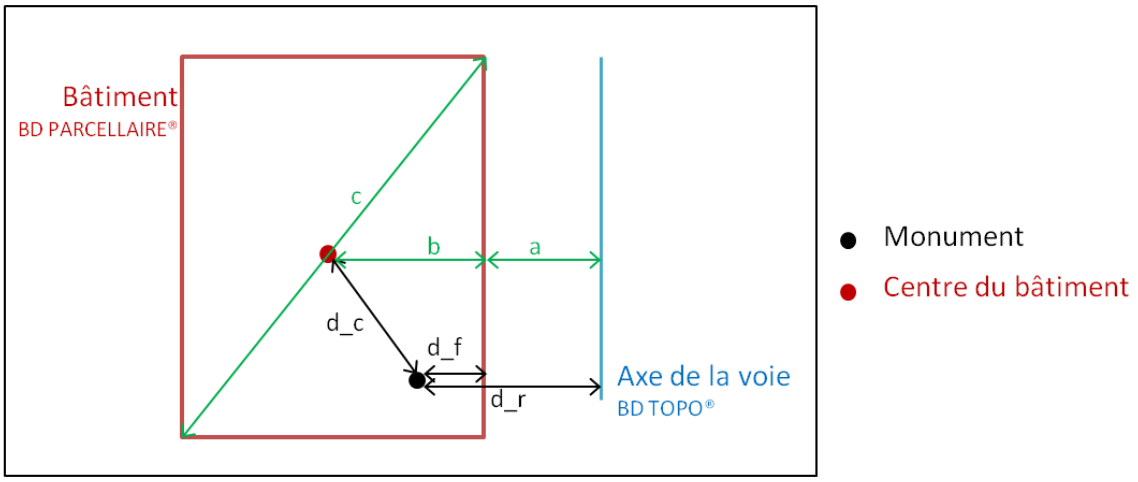


Figure 3.21 Descripteurs spatiaux pour l'apprentissage du modèle de classification des modélisations géométriques des géométries des monuments DBpedia.

Tout comme pour l'exemple précédent, nous avons utilisé Qgis pour le calcul des valeurs des descripteurs pour chaque géométrie de monument. Nous avons également utilisé Weka pour appliquer les algorithmes d'apprentissage supervisé pour obtenir une classification des géométries. La seule différence dans ce cas est que les classes effectives de chaque géométrie sont inconnues. Autrement dit, la fonction $Y: G \rightarrow \{C_1, C_2, C_3\}$ n'est pas définie et l'échantillon d'apprentissage n'est pas fourni. Nous avons donc commencé par préparer cet échantillon en sélectionnant manuellement une trentaine de géométries pour chaque classe. Nous avons ensuite appliqué plusieurs des algorithmes disponibles dans Weka à nos données et nous avons contrôlé manuellement les résultats obtenus. Il s'agit ici des mêmes algorithmes que nous avons utilisés dans le cas des adresses du Grand Lyon qui ont montré leur fiabilité.

Le tableau 3.2 présente les résultats de classification des quatre algorithmes d'apprentissage testés pour assigner à chaque point DBpedia un type de choix de modélisation géométrique. Les quatre algorithmes testés fournissent de bons résultats, ce qui tend à valider notre choix de descripteurs.

Méthode	Précision	Rappel	F-measure
Bayes Network	91,6%	91,3%	91,3%
Jrip	96,3%	96,3%	96,3%
Decision Table	96,4%	96,3%	96,2%
Random Forest	98,8%	98,8%	98,7%

Tableau 3.2 Résultats fournis par Weka pour les quatre algorithmes de classification testés

Enfin, nous avons calculé la précision planimétrique absolue de chaque point DBpedia en additionnant deux valeurs : sa distance avec le point du jeu de données géographiques de référence représentant le choix de modélisation géométrique supposé de son contributeur - barycentre du bâtiment correspondant, façade du bâtiment correspondant ou axe de la rue - et la précision planimétrique absolue du jeu de données géographiques de référence. Ainsi, pour les points classés comme saisis au niveau de l'axe de la rue, la précision planimétrique des données de référence est directement fournie par les valeurs de l'attribut « PREC_PLANI »[95] de chaque segment de route de la BD TOPO® : 1.5 m, 2.5 m, 5 m, 10 m ou 30 m. Pour les points classés comme saisis au centre ou sur la façade d'un bâtiment, c'est la valeur de précision planimétrique moyenne de la BD PARCELLAIRE® qui a été utilisée, soit 10 m[96].

Les résultats de classification et d'estimation de précision planimétrique sont ensuite traduits en données RDF associées à chaque géométrie grâce à des requêtes Sparql d'insertion, de la même manière que pour l'exemple détaillé dans la partie 3.2.1.

Discussion des résultats

Bien que les résultats de classification des références spatiales des monuments DBpedia soient également satisfaisants, ils restent dépendants des hypothèses initiales sur les choix de modélisation géométriques faits par les contributeurs. La définition des types de choix de modélisation géométrique est fondée sur les localisations des ressources DBpedia par rapport aux données géographiques recommandées comme sources de référence par le projet « *WikiProject Geographical Coordinates* ». Or, ceci suppose que nous admettons que le RGE de l'IGN a effectivement été utilisé comme source de données par l'ensemble des contributeurs de Wikipedia. En outre, nous supposons que les coordonnées des ressources DBpedia sont suffisamment précises pour que l'on puisse admettre que le bâtiment le plus proche corresponde bien au monument localisé par ces coordonnées. Ceci suppose encore que les coordonnées retranscrites proviennent d'une source fiable et possèdent 4 décimales (voire 5 pour les plus petits bâtiments) comme recommandé par le projet « *WikiProject Geographical Coordinates* ». Sur les 625 monuments historiques analysés, 606 possèdent des coordonnées dotées d'au moins 4 décimales et 500 des coordonnées dotées d'au moins 5 décimales. Ceci tend à confirmer que les recommandations de saisie de coordonnées sont plutôt respectées sur ce point et que les contributeurs ont eu à cœur de fournir des informations de géolocalisation précises. En revanche, la répartition des ressources dans les trois types de choix de

[95] http://professionnels.ign.fr/sites/default/files/DC_BDTOPO-2-1.pdf
[96] http://professionnels.ign.fr/sites/default/files/DC_BDPARCELLAIRE_1-2.pdf

modélisation géométrique tend à montrer que les recommandations de saisie concernant l'élément de forme caractéristique à représenter ne sont pas suivies. En effet, les ressources sont presque également réparties entre les trois types de choix dans la classification réalisée manuellement : 33.4% pour le centre des bâtiments, 35.4% pour leur façade et 31.2% pour l'axe de la rue. Ceci confirme l'utilité de notre approche d'identification de la modélisation géométrique de chaque géométrie à l'intérieur d'une même source de données.

Les deux exemples que nous avons traités ici concernent des références spatiales ponctuelles. Néanmoins, l'approche que nous avons proposée demeure applicable aux géométries linéaires ou surfaciques à condition de choisir les bons descripteurs d'apprentissage. Bien que le premier exemple des adresses du Grand Lyon soit un cas que nous avons construit nous même à partir de trois modélisation géométrique différentes d'une même base, il s'apparente au cas des monuments DBpedia qui montre un exemple concret d'une source de données dotées de géométries acquises de façon hétérogène, dans lesquelles chaque entité géographique, discernable sur le terrain en tant qu'objet topographique individuel, est localisée à l'aide d'un point saisi au niveau d'un élément caractéristique de sa forme a priori inconnu. En revanche, l'approche est moins adaptée pour les entités géographiques perçues par agrégation d'objets individuels, comme les zones urbaines. (Ruas, 1999) parle d'objets macro pour désigner ces entités géographiques définies comme des populations d'entités géographiques individuelles, elles-mêmes nommées « objets micro ». En effet, il sera plus difficile, dans le cas d'objets « macro » d'identifier le type de choix de représentation effectué, à moins que celui-ci ne vise directement l'un des objets micro le constituant, clairement identifiable au sein de la population des objets micro. Ainsi, elle sera applicable pour une unité administrative si, comme le préconisent les recommandations de saisie du projet « *WikiProject Geographical Coordinates* » par exemple, celle-ci est localisée au niveau du siège de son administration. En revanche, il ne sera pas possible de déterminer de façon précise et consensuelle la localisation du centre de la zone bâtie principale d'une ville pour l'utiliser comme localisation de référence.

La distribution des résultats d'évaluation des précisions planimétriques des géométries des monuments DBpedia est représentée sur la Figure 3.22. On constate que les valeurs des écarts estimées restent majoritairement faibles, avec une forte prédominance des valeurs entre 15 et 25 mètres. Les ressources saisies au niveau des façades ou du centre des bâtiments présentent des valeurs de précision planimétrique supérieures à 10 mètres en raison de la valeur de précision planimétrique moyenne relativement large utilisée pour les bâtiments de la BD PARCELLAIRE®. Celles saisies au niveau de l'axe de la rue présentent des valeurs relativement faibles. Ceci est probablement dû aux indicateurs spatiaux utilisés pour la classification qui exigent des distances à l'axe de la rue plutôt faibles pour assigner une ressource à cette classe, ainsi qu'aux valeurs de précision planimétrique absolue très faibles de la BD TOPO® pour les segments de rues parisiens. Les résultats d'évaluation de la précision planimétrique dépendent bien évidemment des hypothèses initiales de la phase d'identification des choix de modélisation géométrique.

3.3 Conclusion

Dans ce chapitre nous avons présenté un vocabulaire décrivant la sémantique des XY qui permet l'explicitation des caractéristiques des géométries dont les différences d'une ressource à l'autre engendrent des hétérogénéités géométriques. Le vocabulaire prend en compte l'existence éventuelle d'hétérogénéités géométriques internes aux sources de données en permettant de représenter individuellement les caractéristiques de chaque géométrie. Le vocabulaire s'appuie sur

les normes et travaux existants en matière de représentation de la qualité et des spécifications de saisie des données géographiques. Le vocabulaire est exprimé en OWL et publié[97] selon les bonnes pratiques du Web de données (W3C, 2008).

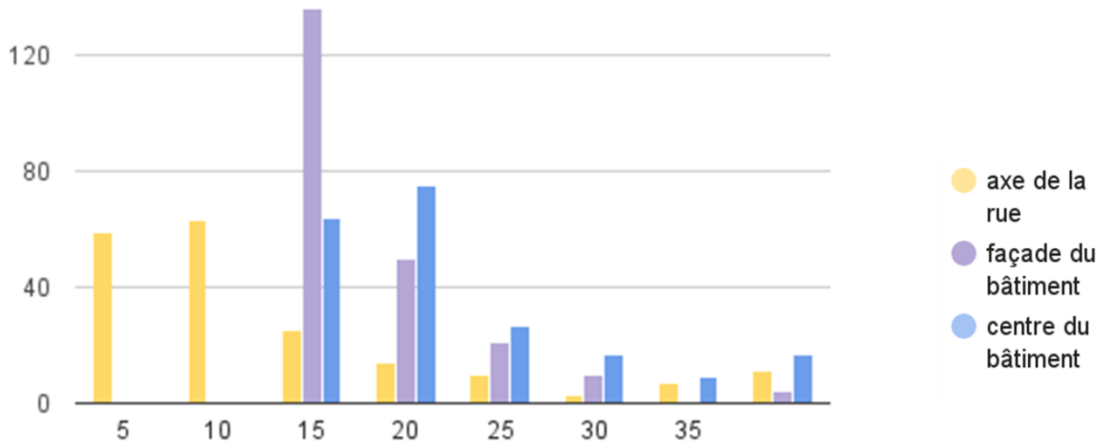


Figure 3.22 Fréquence des valeurs de précision planimétrique pour les monuments historiques de Paris de DBpedia en fonction du type de choix de modélisation. L'axe des abscisses indique les valeurs de précisions maximales, en mètres et celui des ordonnées le nombre de ressources DBpedia.

En deuxième partie nous avons proposé différentes solutions pour peupler ce vocabulaire. Nous avons distingué les deux cas de figure possibles, à savoir la présence ou l'absence de métadonnées de qualité fournies avec les données. Dans le premier cas, nous avons démontré comment on peut extraire et traduire les métadonnées liées aux caractéristiques des géométries en RDF pour les associer aux ressources géométriques correspondantes. Le second cas soulève plus de défis. Pour cela, nous avons proposé une approche basée sur l'utilisation des données géographiques de référence qui permet d'identifier automatiquement, grâce aux méthodes de classification par apprentissage supervisé, le choix de modélisation géométrique de chaque géométrie d'une source de données. Cette méthode permet également de déduire une estimation de la précision planimétrique de chaque géométrie en utilisant la méthode d'évaluation directe classiquement utilisée dans le domaine des données géographiques. Cette approche reste quand même basée sur des hypothèses fortes sur les intentions des créateurs des données en matière de choix de modélisation géométrique. Elle dépend également de la nature des entités géoréférencées et de la disponibilité de données géographiques de référence sémantiquement compatibles avec cette nature. Même si les résultats de classification obtenus par cette approche peuvent être hypothétiques, ils permettent de fournir une qualification individuelle des géométries que nous considérons nécessaires dans la prise ne compte de l'hétérogénéité géométrique dans un processus d'interconnexion. Nous présentons dans la suite comment un processus d'interconnexion qui utilise la comparaison géométrique comme critère principal de mise en correspondance, peut adapter cette comparaison selon les caractéristiques des géométries à traiter.

[97] http://data.ign.fr/def/xysemantics/

4 PROPOSITIONS D'APPROCHES D'INTERCONNEXION ET DE VISUALISATION DES DONNÉES GÉORÉFÉRENCÉES SUR LE web DE DONNÉES

Dans ce chapitre, nous présentons deux approches pour améliorer la prise en compte d'un critère d'interconnexion spatial entre données géoréférencées du Web présentant une forte hétérogénéité au niveau des géométries y compris au sein d'un même jeu de données. La première s'appuie sur des connaissances explicites sur les caractéristiques des géométries pour adapter dynamiquement les paramètres du processus d'interconnexions aux cas à traiter. La seconde s'appuie sur des données géographiques de référence comme support pour l'interconnexion afin de surmonter les hétérogénéités géométriques entre les références spatiales des ressources du Web. Les liens créés entre données thématiques et référentiel géographique peuvent être mis à profit pour faciliter savoir la visualisation cartographique des données à différentes échelles. Nous présenterons donc différentes solutions de visualisation multi-échelle de données liée mettant en œuvre les données géographiques de références et les liens de correspondance.

4.1 Une approche à base de connaissances pour adapter dynamiquement le paramétrage du processus d'interconnexion de données géoréférencées

Nous avons vu dans le chapitre 2, notamment dans la section 2.2, que la configuration d'un processus d'interconnexion de données consiste à définir, d'une manière manuelle ou automatique, les différents paramètres du processus. Dans les outils d'interconnexion génériques tels que Silk ou LIMES, ceci consiste à spécifier les règles d'interconnexion qui précisent principalement le choix des propriétés à comparer, les mesures de distance à utiliser pour comparer ces propriétés, les seuils d'acceptation de chaque mesure et éventuellement la fonction d'agrégation ainsi que le poids attribué à chaque mesure dans le cas d'une comparaison multicritère. Fixer ces choix manuellement nécessite l'intervention d'un expert des données qui comprend les hétérogénéités entre les sources de données et peut donc choisir les propriétés à comparer ainsi que les distances, seuils et poids adéquats. Comme nous l'avons vu dans le chapitre 2, de nombreuses approches de l'état de l'art proposent des méthodes qui permettent d'automatiser, ou au moins de faciliter cette tâche de configuration pour l'expert des données. Cependant, très peu d'approches se sont concentrées sur la prise en compte des caractéristiques des valeurs des propriétés, plus particulièrement celles des références spatiales, dans la configuration des différents paramètres d'interconnexion. Nous présentons ici une approche qui s'appuie sur les caractéristiques des géométries représentées conformément au vocabulaire de la sémantique des XY présenté dans le chapitre 3 pour adapter dynamiquement les paramètres de comparaison de chaque paire de références spatiales. Nous présentons d'abord une formalisation de cette approche afin de l'inclure dans un contexte d'interconnexion multicritère. Nous fournissons ensuite une description générale de l'approche d'adaptation dynamique de la comparaison des géométries. Nous expliquons les choix de mise en œuvre de cette approche puis finalement les tests effectués pour sa validation.

4.1.1 Formalisation de l'approche

Dans un souci de généricité, nous souhaitons que notre approche puisse s'inscrire dans un processus d'interconnexion multicritère, tel que ceux mis en œuvre dans des outils comme Silk ou LIMES. L'interconnexion multicritère peut être perçue comme un problème classique de décision multicritère. L'interconnexion peut être formalisées comme une opération qui prend en entrée un deux jeux de données, un jeu de données source S et un jeu de données cible T et qui produit un ensemble de liens entre les entités des deux jeux de données. Ces liens représentent des relations de correspondances entre entités sémantiquement liées (Ferrara et al., 2013). Pour chaque ressource $s \in S$ le but est donc de décider avec quelle ressource $t_i \in T$ elle doit être interconnectée. Supposons le cas le plus simple où une ressource $s \in S$ a deux candidats à l'interconnexion $t_1, t_2 \in T$, tels que a est le lien possible à créer entre s et t_1 et b le lien possible entre s et t_2. Il s'agit de décider lequel des deux liens a et b doit être créé, c.-à-d. établir un ordre de préférence entre a et b.

Il s'agit donc bien d'un problème de décision multicritère, car on peut comparer a et b sur différents critères. Supposons qu'on dispose de n critères de comparaison pour l'interconnexion. Deux stratégies principales sont souvent appliquées pour établir un ordre de préférence: comparer puis agréger ou agréger puis comparer (Grabisch et Perny, 2003). La première stratégie, dite de surclassement, consiste à comparer les candidats (les liens candidats dans notre cas) critère par critère pour établir des ordres de préférence partiels (un ordre de préférence par critère). Ensuite ces ordres de préférence sont agrégés pour établir un ordre général entre les candidats. La deuxième stratégie consiste à agréger les scores obtenus par les différents critères pour chaque candidat à classer (lien candidat dans notre cas), puis comparer les scores agrégés des différents candidats. Cette technique est la plus souvent utilisée dans les outils d'interconnexion multicritère tels que LIMES et Silk. Entre les liens candidats a et b, l'ordre de préférence est déterminé dans ce cas selon l'équation 4.1.

$$(g_1(a), \ \ldots, g_n(a)), (g_1(b), \ \ldots, g_n(b)) \xrightarrow{\psi} v(a), v(b) \xrightarrow{\Phi} \gtrsim (a, b) \qquad 4.1$$

Où $\gtrsim$ représente la relation de préférence. $g_i(x)$ est la fonction d'évaluation du *i*ème critère pour le lien candidat x. ψ est une fonction d'agrégation et Φ est la fonction de comparaison.

Dans un processus d'interconnexion, la valeur retournée par la fonction g_i n'est autre que la similarité calculée pour l'*i*ème critère entre s et t_1 pour le lien candidat a et entre s et t_2 pour le lien candidat b. $v(x)$ représente la fonction de valeur de similarité globale associée au lien candidat x. Les fonctions d'évaluation des critères g_i ainsi que la fonction d'agrégation ψ sont à fixer en amont pour l'interconnexion.

Notre proposition consiste donc à adapter ces fonctions g_i et ψ pour le critère spatial en fonction des caractéristiques des géométries des ressources s, t_1 et t_2. Ainsi, pour chaque comparaison entre deux ressources, le calcul de la valeur de similarité des références spatiales ainsi que sa prise en compte dans la fonction d'agrégation doivent être adaptés aux caractéristiques des références spatiales de ces ressources. Supposons que le critère spatial soit le *k*ième critère de comparaison entre les ressources. Nous proposons d'adapter les fonctions g_k et ψ aux caractéristiques des références spatiales des ressources liées par les liens candidats comme suit :

$$\left.\begin{array}{l}(\,g_1(a),\ldots,\,g_{ak}(a),\ldots,g_n(a)) \xrightarrow{\psi_a} v\,(a)\\ (g_1(b),\ldots,\,g_{bk}(b),\ldots,g_n(b)) \xrightarrow{\psi_b} v\,(b)\end{array}\right\} \xrightarrow{\Phi} \gtrsim (a,b) \qquad 4.2$$

4.1.2 Description générale de l'approche

La configuration des différents paramètres d'interconnexion est une tâche qui peut s'avérer compliquée : il faut impérativement connaître les données et leurs hétérogénéités pour pouvoir choisir les bons paramètres. Comparer les références spatiales plus précisément nécessite des connaissances sur leurs caractéristiques. Par exemple, fixer le seuil d'une distance entre deux géométries nécessite de disposer de connaissances permettant d'évaluer l'écart maximal que peuvent présenter deux géométries représentant un même phénomène spatial. Or, dans un contexte où les caractéristiques des géométries sont hétérogènes à l'intérieur d'une même source, fixer ce paramètre à une seule valeur peut être pénalisant : dans certains cas, cela peut engendrer des liens erronés (des faux positifs), dans d'autres cas cela peut faire omettre de vrais liens (des faux négatifs). Nous souhaitons adapter dynamiquement le paramétrage de l'interconnexion à un niveau plus fin, à savoir au niveau de chaque comparaison géométrique. Cette adaptation dynamique de la configuration de la comparaison géométrique doit être indifféremment utilisable dans une interconnexion monocritère ou dans un processus d'interconnexion multicritère plus générique. Les paramètres de comparaison des géométries que nous proposons d'adapter sont:

- Le choix de la mesure de distance
- Le comportement de la fonction de similarité
- La valeur du seuil de distance
- Le poids associé à la valeur de similarité calculée pour des géométries dans le cas d'une interconnexion multicritère avec une fonction d'agrégation pondérée
- La neutralité du critère géométrique dans une interconnexion multicritère : on se réserve la possibilité de ne pas prendre en compte ce critère quand les caractéristiques des géométries le justifient

Dans les outils d'interconnexion classiques, ces paramètres sont souvent choisis par l'utilisateur qui configure l'interconnexion selon son expertise sur les données en suivant un raisonnement qui peut se présenter le plus souvent sous forme de règles de décision. Nous expliquons dans la suite comment ceci est concrétisé pour chacun de ces paramètres. Notons que les règles présentées ici pour chaque paramètre sont proposées à titre d'exemple et ne sont donc ni exhaustives ni universelles.

Choix de la mesure de distance

Le choix de la mesure de distance revient à décider laquelle est la mieux adaptée pour comparer les références spatiales. Nous avons vu dans la section 2.1.1 les mesures de distance qui correspondent aux différentes primitives géométriques. L'expert chargé de l'interconnexion précise celle qui est la plus adaptée au cas d'application. Par exemple, s'il s'agit de données géoréférencées par des points dotés de coordonnées planes dans les deux jeux de données, le choix le plus évident est d'utiliser une distance euclidienne pour comparer les références spatiales. Si les ressources sont géoréférencées par des polygones dans les deux jeux de données, l'expert peut, par exemple, choisir

d'utiliser la mesure de recouvrement surfacique entre les surfaces des polygones. Les règles que l'expert suit pour établir son choix peuvent être formulées comme présenté dans l'Encadré 4.1.

> **SI** *les géométries des deux jeux de données sont des points* **ALORS** *utiliser la mesure de distance euclidienne.*
>
> **SI** *les géométries des deux jeux de données sont des polygones* **ALORS** *utiliser la mesure de distance de recouvrement surfacique.*
>
> **…**

Encadré 4.1 Quelque exemple de choix de mesures de distance entre géométries

Comportement de la fonction de similarité

La valeur de similarité entre deux valeurs de propriétés est souvent définie en fonction de la valeur de distance entre celles-ci. Dans les outils Silk et LIMES, cette fonction est figé et l'expert ne peut pas facilement intervenir sur sa définition. Dans le cas de LIMES, plusieurs mesures permettent de calculer directement la similarité entre deux valeurs sans avoir recours à un calcul de distance préalable (*ex.* mesure de similarité de *Jaccard*). Néanmoins, en cas de comparaison des géométries, la similarité est calculée à partir de la valeur de distance selon l'équation 4.3.

$$similarité = \frac{1}{1 + distance} \tag{4.3}$$

Cette fonction a donc un comportement fixe : elle est décroissante par rapport à la distance et permet de fournir une valeur normalisée qui appartient à l'intervalle $]0, 1]$ (voir Figure 4.1).

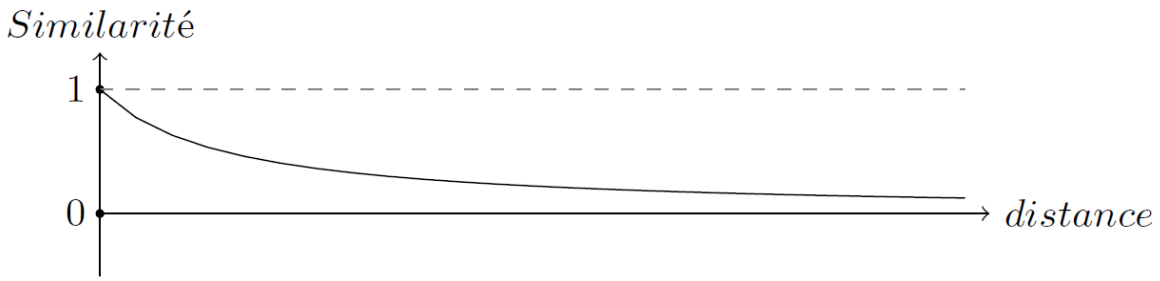


Figure 4.1 Calcul de similarité à partir de la valeur de distance spatiale dans LIMES

Dans le cas de Silk, toutes les mesures implémentées fournissent une valeur de distance. Cette valeur est ensuite convertie en une valeur de similarité, appelée valeur de confiance[98], pour chaque critère selon l'équation 4.4, telle que θ représente le seuil de distance.

[98] https://github.com/silk-framework/silk/blob/develop/doc/LinkageRules.md#parameters

$$similarité = \begin{cases} 1 - \dfrac{distance}{\theta} \; si \; distance \in [0, 2.\theta] \\ -1 \; sinon \end{cases} \quad 4.4$$

L'intervention de l'expert dans la définition de la fonction de similarité est limitée au choix du seuil de distance θ. Le comportement de la fonction demeure le même : c'est une fonction linéaire décroissante de valeurs normalisées appartenant à l'intervalle $[-1, 1]$ et qui s'annule quand la distance est égale au seuil (Figure 4.2). La particularité de la fonction de similarité ici est sa capacité à fournir des valeurs négatives. La valeur de similarité calculée pour un critère peut donc avoir un effet négatif sur la valeur de similarité globale.

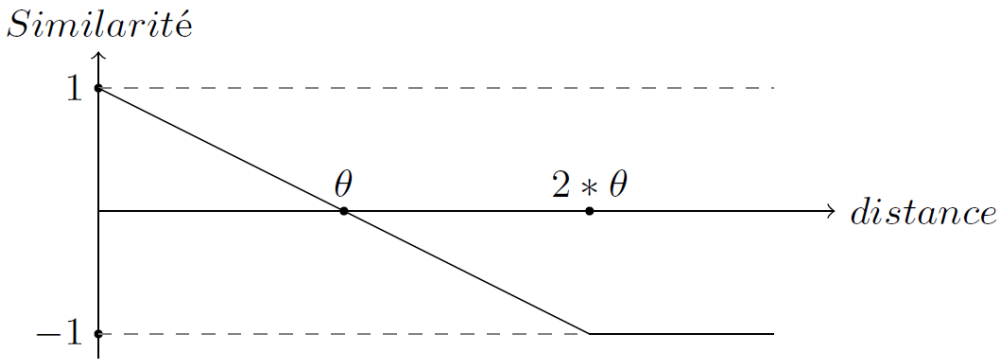


Figure 4.2 Calcul de similarité (confiance) dans Silk à partir de la valeur de distance d et du seuil Θ.

Choix du seuil de la mesure de distance

Dans un processus d'interconnexion, choisir un seuil de distance pour un critère donné revient à fixer une limite de distance au-delà de laquelle des valeurs des propriétés comparées sont considérées comme non similaires (distinctes). Dans le cas de LIMES, les mesures implémentées fournissent en sortie une valeur de similarité, qu'elle soit calculé à partir d'une mesure de distance ou non. Le seuil fourni par l'utilisateur pour chaque critère est donc un seuil de similarité qui précise la limite de au-dessous de laquelle les deux valeurs comparées sont considérées comme non similaires. Dans le cas de Silk, comme nous l'avons vu avant, le choix de la valeur du seuil de distance est essentiel au calcul de la similarité. La définition d'un seuil de distance peut être très délicate. L'expert qui configure l'interconnexion doit s'assurer de choisir une valeur suffisamment large pour associer les valeurs homologues, et suffisamment petite pour éviter l'association de valeurs distinctes.

En ce qui concerne la comparaison des géométries, l'expert de l'interconnexion fixe le seuil de distance par estimation de l'écart maximal possible entre les géométries des ressources en s'appuyant le plus souvent sur les connaissances qu'il possède sur celles-ci. Par exemple, s'il s'agit de géométries qui sont supposées être localisées au même endroit pour les ressources homologues, l'expert peut estimer le seuil en prenant en compte l'écart possible entre les géométries dû à leurs imprécisions. Le seuil peut donc être défini comme à la somme des précisions planimétriques des deux jeux de données, en considérant qu'au-delà de cette somme les géométries ne peuvent pas représenter la même entité géographique. Dans le cas où les géométries des ressources sont modélisées différemment entre les deux sources (*ex.* ressources localisées par des points sur l'axe de la route dans le premier jeu de données et ressources localisées par des points sur le bord de la

route), l'expert peut fixer le seuil à la somme des précisions planimétriques des jeux de données auquel s'ajoute l'écart possible entre les deux modélisations géométriques (*ex.* l'écart moyen entre l'axe et le bord d'une route dans la zone qui contient les données). Le raisonnement que l'expert suit dans ce cas peut également être formulées selon des règles comme celles présentées dans l'Encadrée 4.2.

SI *les géométries sont modélisées de la même manière dans les deux jeux de données et la précision planimétrique des géométries du premier jeu est égale à* P_1 *et la précision planimétrique des géométries du deuxième jeu est égale à* P_2 **ALORS** *le seuil est fixé à* $P_1 + P_2$.

SI *les géométries sont modélisées de manières différentes dans les deux jeux de données et la précision planimétrique des géométries du premier jeu est égale à* P_1 *et la précision planimétrique des géométries du deuxième jeu est égale à* P_2 *et l'écart estimé entre les deux modélisations géométriques est égal à E* **ALORS** *le seuil est fixé à* $P_1 + P_2 + E$.

…

Encadré 4.2 Quelque exemple de règles pour définir le seuil de la disatnce géométrique

Choix du poids de la similarité des géométries dans une agrégation multicritère

Dans le cas où le processus d'interconnexion met en œuvre une agrégation multicritère pondérée, comme une moyenne arithmétique pondérée des valeurs de similarité des différents critères pour le calcul de la similarité globale, les poids des différents critères sont fixés au préalable. Le poids de chaque critère représente l'importance que l'on souhaite lui attribuer dans le calcul de la similarité globale dont la valeur servira à décider de l'opportunité de créer un lien ou pas. Dans le cas de LIMES, les opérateurs de combinaisons des valeurs de similarités calculées pour les critères de comparaison sont limités aux opérations logiques[99] « ET », « OU » et « DIFF », et les opérations arithmétiques[100] « MAX », « MIN » et « ADD ». Il n'existe donc aucune fonction d'agrégation pondérée pour les valeurs de similarités dans LIMES. En revanche dans Silk, en plus des opérateurs d'agrégation qui renvoie le « Minimum », le « Maximum » et la « Moyenne quadratique », il existe deux opérateurs d'agrégation pondérée: la « Moyenne arithmétique » et la « Moyenne géométrique »[101].

Fixer le poids d'un critère de comparaison dans une interconnexion multicritère dépend le plus souvent de la qualité des valeurs comparées pour ce critère. Pour fixer le poids du critère de comparaison des géométries, l'expert peut estimer par exemple que le poids dépend de la qualité des géométries en termes de précision planimétrique. Il peut donc, selon le contexte et la nature des données, décider d'accorder un poids important s'il estime que les deux jeux de données sont de bonne précision planimétrique (*ex.* une précision planimétrique inférieure à 10m pour les deux jeux de données). Sinon il lui accorde un poids moins important que ceux attribués aux critères de comparaison considérés comme plus fiables. Quand l'expert sait que les entités représentées dans l'un des deux jeux de données ont des contours vagues et sont donc difficiles à localiser de façon

[99] https://github.com/AKSW/LIMES-dev/blob/master/limes-core/manual/user_manual/configuration_file/metric/boolean_operations.md
[100] https://github.com/AKSW/LIMES-dev/blob/master/limes-core/manual/user_manual/configuration_file/metric/metric_operations.md
[101] https://github.com/silk-framework/silk/blob/develop/doc/LinkageRules.md#aggregations

consensuelle par des géométries vectorielles, il peut décider de fixer le poids du critère géométrique au minimum. Ces décisions peuvent être formulées selon les règles présentées dans l'Encadré 4.3.

SI *la précision planimétrique des géométries des deux jeux de données est inférieure à 10m et les poids des autres critères son a et b* **ALORS** *le poids est fixé à une valeur c> a et b.*

SI *la précision planimétrique des géométries de l'un des deux jeux de données est supérieure à 10m et les poids des autres critères son a et b* **ALORS** *le poids est fixé à une valeur c< a et b.*

SI *les ressources des deux jeux de données représentent des entités géographiques vagues et les poids des autres critères son a et b* **ALORS** *le poids est fixé à une valeur c< a et b.*

…

Encadré 4.3 Quelque exemple de règles pour définir le poids du critère géométrique

Neutralité du critère de comparaison des géométries

Décider de la neutralité d'un critère d'appariement revient à spécifier s'il faut l'inclure ou pas dans une interconnexion. Il s'agit principalement d'une décision prise par l'expert de l'interconnexion en amont plus que d'un paramètre à fixer. L'expert peut se référer à la qualité des données pour décider si un critère de comparaison possible est utilisable ou pas pour l'interconnexion. Par exemple si l'expert considère qu'entre les deux jeux de données, la comparaison des labels des ressources est un critère d'interconnexion possible, mais après l'analyse des données, les valeurs de ces labels s'avèrent très hétérogènes ou peu renseignées, l'expert peut décider de ne pas utiliser ce critère pour l'interconnexion. Dans le cas de la comparaison des géométries, la décision peut par exemple être basée sur la qualité de précision planimétrique des géométries. L'expert peut considérer, pour un contexte précis, que le critère de comparaison de géométries ne doit pas être pris en compte dans l'interconnexion si la précision planimétrique de l'un des deux jeux de données dépasse les un certain seuil. La possibilité de renoncer à utiliser la comparaison des géométries pour l'interconnexion est particulièrement utile dans le cas de données représentant des entités géographiques dotées de contours vagues. Dans ce cas, les décisions peuvent être formulées selon les règles présentées dans l'Encadré 4.4.

SI *les ressources des deux jeux de données représentent des entités géographiques vagues* **ALORS** *ne pas utiliser le critère de comparaison des géométries dans l'interconnexion.*

SI *la précision planimétrique des géométries de l'un des deux jeux de données est supérieure à un x mètres* **ALORS** *ne pas utiliser le critère de comparaison des géométries dans l'interconnexion.*

…

Encadré 4.4 Quelque exemple de règles pour décider de la neutralité du critère de comparaison géométrique

Adaptation des paramètres de l'interconnexion

À travers les exemples précédents, nous notons que l'intervention de l'expert pour choisir les différents paramètres pour le critère de comparaison des géométries se fait d'une manière globale pour les jeux de données. L'expert se réfère à des caractéristiques globales des sources de données,

comme la précision planimétrique globale ou la modélisation géométrique globale, qui supposent une homogénéité à l'intérieure de chaque source. Dans le cas d'une hétérogénéité des géométries à l'intérieur des sources de données, fixer les paramètres d'interconnexion d'une manière globale devient une tâche très difficile. Il devient même impossible de trouver le meilleur compromis pour fixer les paramètres dans certains cas.

Notre proposition consiste à pouvoir varier ces différents paramètres de comparaison des géométries pour mieux les ajuster aux caractéristiques de chaque paire de références spatiales comparées. Dans ce cas, fixer un paramètre à une valeur précise ne doit pas être décidé globalement, mais localement au niveau de chaque comparaison.

Afin de permettre la définition des différents paramètres de configuration du processus d'interconnexion, nous proposons de formaliser et utiliser des règles de décision qui prennent en entrée les caractéristiques des deux références spatiales à comparer et qui donnent en sortie les valeurs des différents paramètres de comparaison (voir Figure 4.3). La définition et la formalisation des règles de décision doivent donc être réalisées en amont par un expert en interconnexion de données géographiques.

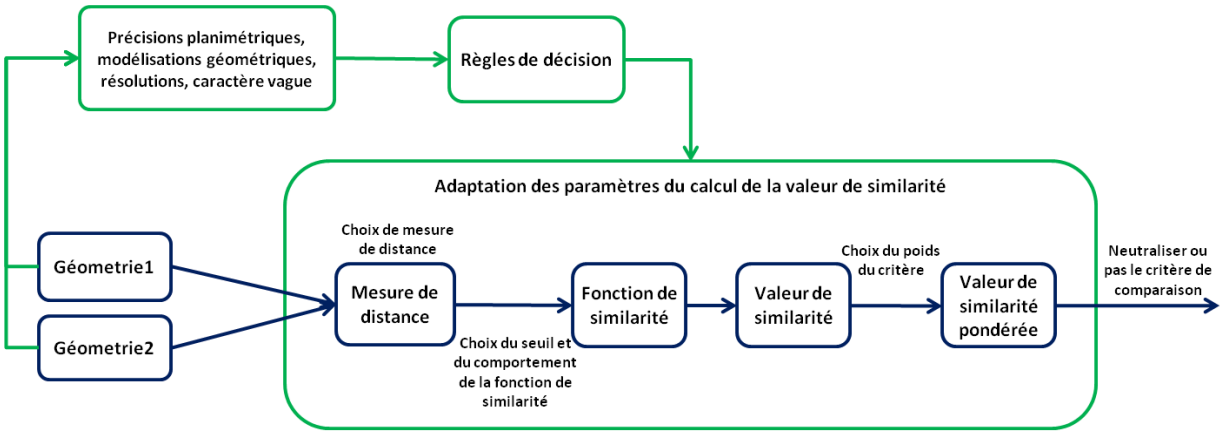


Figure 4.3 L'approche d'adaptation dynamique du paramétrage des comparaisons des références spatiales en fonction de leurs caractéristiques

Le comportement de la fonction de similarité de Silk présentée par l'équation 4.4 est rigoureusement le même pour chaque critère d'interconnexion et pour toute paire de ressources comparées. Nous proposons d'adapter cette fonction pour la comparaison des références spatiales comme suit :

$$similarité = \begin{cases} 1 - \left(\frac{d}{\theta}\right)^{\alpha} & si\ d \in [0, \theta] \\ -\left(\frac{d-\theta}{\theta}\right)^{\beta} & si\ d \in [\theta, 2.\theta] \\ -1\ sinon \end{cases} \quad 4.5$$

Tel que d représente la distance calculée entre les deux références spatiales par la mesure de distance choisie par les règles de décision en plus des paramètres θ, α et β. Le paramètre θ représente le seuil. Il peut être défini par exemple à partir des précisions planimétriques des références spatiales. Les paramètres α et β représentent des facteurs de convexité et de concavité des parties positive et négative de la fonction de similarité. α et β doivent être des valeurs positives. La fonction de similarité devient convexe lorsque ces paramètres prennent des valeurs supérieures à

1 et concave lorsqu'ils prennent des valeurs inférieures à 1 (voir Figure 4.4). Le changement des paramètres θ, α et β par les règles de décision permet de rendre la fonction de similarité plus ou moins stricte tout en gardant sa monotonie. Deux autres paramètres sont éventuellement nécessaires : le premier est ω qui représente le poids de la valeur de similarité géométrique dans le cas d'une agrégation multicritère. Le deuxième est le paramètre de neutralité λ qui peut porter une valeur booléenne qui permet de décider de la prise en compte ou pas du critère de comparaison géométrique.

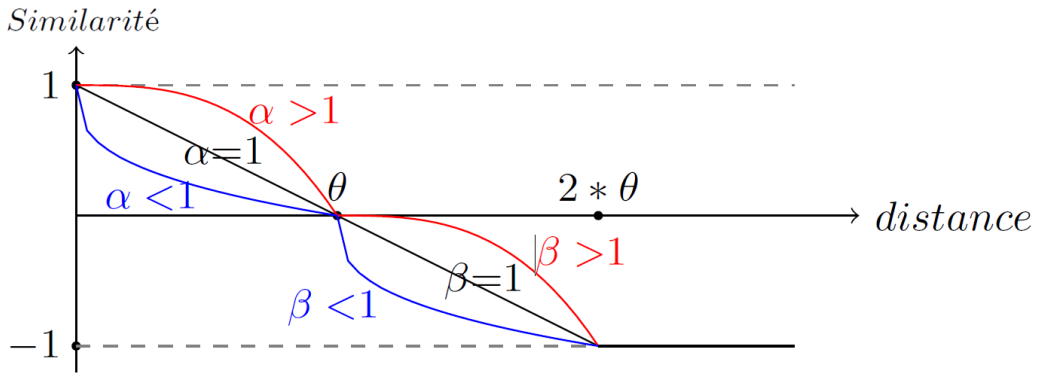


Figure 4.4 Variation du comportement de la fonction de similarité

4.1.3 Mise en œuvre de l'approche d'adaptation dynamique des paramètres de la comparaison des géométries.

Afin de tirer profit des avantages des outils d'interconnexion génériques, nous avons étudié la possibilité d'implémenter notre approche dans les outils Silk et LIMES. Nous avons choisi d'adapter Silk[102] en rendant la comparaison des géométries compatible avec l'approche d'adaptation dynamique présentée avant. Nous présenterons dans la suite les principes de fonctionnement généraux de Silk qui nous ont incités à le choisir pour l'implémentation.

Principes de fonctionnement et architecture générale de Silk

Nous avons choisi l'outil Silk comme base pour l'implémentation de notre approche, car il représente, avec LIMES, les outils les plus génériques et les plus utilisés pour l'interconnexion. LIMES n'étant pas ouvert au moment où nous avons commencé l'implémentation de notre approche, nous nous sommes naturellement orientés vers l'utilisation de Silk. Silk présente une multitude d'avantages tels que la variété des opérateurs de transformation, de comparaison de données et d'agrégation de scores de similarité déjà implémentés ainsi que la possibilité de rajouter de nouveaux opérateurs de transformation ou de mesures de similarité sous forme de plugins. Silk implémente également des méthodes d'indexation et de partitionnement des données afin de réduire le nombre de comparaisons à effectuer et optimiser donc le temps d'exécution d'une interconnexion.

---

[102] Nous avons utilisé la version 2.6 de Silk qui était la dernière au moment où on a commencé l'implémentation.

Silk est écrit dans le langage Scala. Trois variantes principales sont disponibles pour Silk: Silk Workbench qui est une version fournie avec une interface graphique, Silk SingleMachine qui est une version fournie sous forme d'un fichier exécutable destiné à être utilisé sur une seule machine et Silk MapReduce qui est une version implémentée grâce à la bibliothèque Hadoop qui permet le traitement distribué des tâches de chargement et d'interconnexion pour le passage à l'échelle. Ces trois versions partagent le même noyau. Toutefois, nous nous sommes concentrés sur la version SingleMachine car elle est la plus stable des trois versions.

Un processus d'interconnexion dans Silk est composé de cinq étapes principales. Ces étapes, ainsi que les classes principales responsables de leur déroulement sont représentées dans la Figure4.5. Silk est conçu selon une architecture à base de tâches. Chaque étape présentée dans la figure est réalisée pendant l'interconnexion sous forme d'une seule tâche ou de plusieurs tâches parallélisées quand c'est possible.

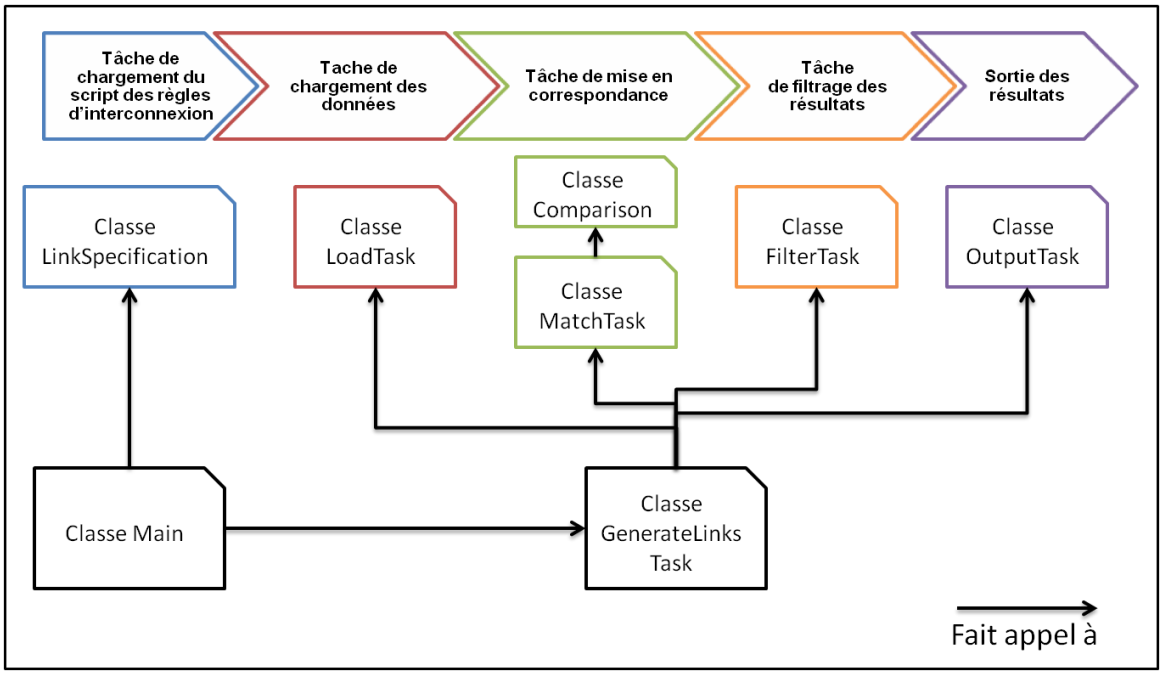


Figure 4.5 Étapes générales et classes principales dans la version SingleMachine de Silk 2.6

Comme nous l'avons expliqué dans la section 2.2.2., Silk est un outil qui nécessite une configuration initiale sous forme d'un fichier de spécifications des règles d'interconnexion. En effet, Silk définit son propre langage déclaratif, dans une syntaxe XML, appelé Silk-LSL[103] (*Link Specification Language*). Un fichier de configuration défini dans ce langage permet de spécifier les jeux de données source et cible et de sélectionner les ressources à interconnecter à partir de ces jeux de données. Il permet également de définir les règles de mise en correspondance qui permettent la sélection : des propriétés à comparer, des opérateurs de transformation éventuels des valeurs de ces propriétés, des mesures de distance pour comparer ces valeurs et des fonctions d'agrégation de scores de confiance calculés par les mesures de distance dans le cas d'une comparaison multicritère. Ce

[103]https://github.com/silk-framework/silk/blob/develop/doc/LinkSpecificationLanguage.md#silk-link-specificsation-language

langage permet également de spécifier les règles de filtrage des liens résultants et de leur écriture en sortie.

Lors du lancement de Silk, la classe principale commence par créer une instance de la classe *LinkSpecification*, qui consiste à charger les spécifications qui indiquent : les types de ressources à charger, les chemins des propriétés à comparer à partir de ces ressources, les mesures de distance à utiliser, etc. Cette instance de la classe *LinkSpecification* est ensuite passée à la tâche générale qui gère le reste du processus, qui est une instance de la classe *GenerateLinksTask* (voir Figure 4.5).

La tâche générale fait appel dans premier temps à la tâche de chargement qui une instance de la classe *LoadTask*. C'est cette tâche qui se charge de récupérer les ressources et les valeurs des propriétés, indiquées par les spécifications d'interconnexion, à partir des sources de données et de les charger dans un cache utilisable pour la tâche suivante. C'est lors de cette tâche de chargement que l'indexation des valeurs et le partitionnement de données ont lieu pour alléger la tâche suivante. En effet, chaque mesure de distance est définie avec une fonction d'indexation qui permet d'affecter, lors du chargement des données, les valeurs proches des propriétés à comparer aux mêmes blocs. Par exemple, si une mesure de distance fondée sur une différence numérique est déclarée dans les spécifications d'interconnexion entre deux propriétés *a* et *b*, cette mesure de distance demande optionnellement de renseigner le maximum et le minimum des valeurs possibles pour les propriétés *a* et *b* (ex. voir Figure 4.6). L'intervalle entre la valeur maximale et la valeur minimale est divisé sur le nombre de blocs spécifié dans les spécifications d'interconnexion afin de calculer la taille de chaque bloc. Si cette taille est inférieure au seuil, elle est remplacée par le seuil. Lors du chargement des données, chaque ressource est affectée au(x) bloc(s) dans lesquels la valeur de sa propriété (*a* ou *b*) est située. Dans Silk, chaque catégorie de mesures de distance emploie une méthode d'indexation spécifique.

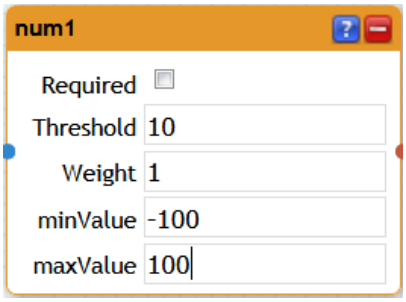


Figure 4.6 Exemple de paramétrage de mesure de distance numérique dans l'interface graphique de Silk

Dès que la tâche de chargement est terminée, une tâche de mise en correspondance est lancée en instanciant la classe *MatchTask*. Cette tâche parcourt les jeux de données source et cible puis applique les règles d'interconnexion, décrites dans les spécifications, sur chaque paire de ressources à comparer. En fait, ces paires sont sélectionnées à partir des ressources qui appartiennent aux mêmes blocs d'indexation. En effet, le partitionnement des données dans des blocs permet se diviser cette tâche de mise en correspondance en plusieurs sous-tâches qui s'exécutent parallèlement (selon la capacité de la machine) pour réduire le temps d'exécution. Le travail de chaque tâche de mise en correspondance consiste donc à parcourir, d'une manière séquentielle, les ressources des jeux de données source et destination appartenant à un bloc pour effectuer des comparaisons. L'application des règles d'interconnexion sur une paire de ressources revient à faire appel à la classe *Comparison* pour chaque paire de propriétés à comparer et éventuellement à la classe *Aggregation* dans le cas

d'une interconnexion multicritère. C'est la classe *Comparison* qui se charge de faire appel à la mesure de distance pour calculer la distance puis la transformer en une valeur de confiance conformément à l'Équation 4.4. Dans le cas d'une agrégation multicritère, la valeur de confiance est envoyée avec son poids de la classe *Comparison* à la classe *Aggregation*.

Une fois la (les) tâches (sous-tâches) de mise en correspondance terminée(s), Silk lance une tâche de filtrage des liens résultants, en ne gardant que les liens ayant un score de similarité (confiance) strictement positive, et qui respectent les cardinalités des liens choisies dans la configuration. La tâche finale consiste à écrire les résultats en sortie dans un fichier ou dans un triple store. Les liens à écrire peuvent être également filtrés : on peut éventuellement spécifier la valeur maximale ou minimale de confiance qu'un lien doit avoir pour être écrit dans un fichier en sortie. Plusieurs fichiers de sortie peuvent être spécifiés.

Implémentation de l'approche dans Silk

La Figure 4.7 représente l'architecture générale l'implémentation de notre approche dans Silk qui s'exécute au niveau de la comparaison des géométries vectorielles.

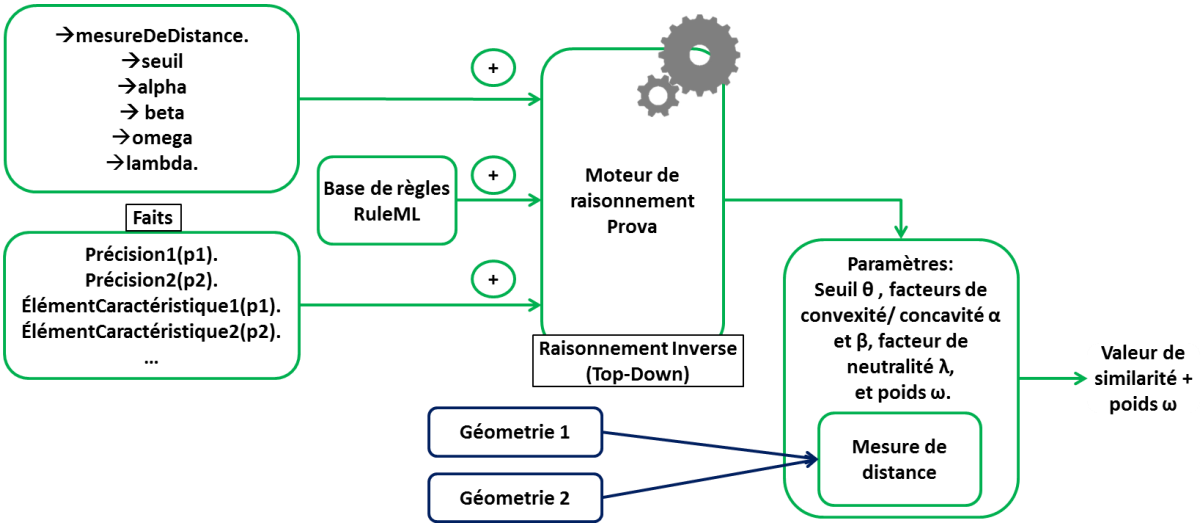


Figure 4.7 Implémentation de l'approche d'adaptation dynamique des paramètres d'interconnexion

Nous avons effectué quatre modifications principales dans Silk pour que cette implémentation puisse fonctionner. La première modification porte sur la classe *LinkSpecification* (voir Figure 4.5). Nous avons modifié cette classe en rajoutant les chemins vers les caractéristiques des géométries à la liste des propriétés dont les valeurs doivent être extraites pour chaque ressource. Le chargement de ces valeurs de métadonnées est conditionné par le fait que la catégorie de la comparaison annoncée dans les spécifications soit « spatiale ». La deuxième modification porte sur le rajout d'une classe d'un moteur de raisonnement. La troisième consiste à instancier la classe de moteur de raisonnement au niveau des instances la classe *MatchTask*. Nous expliquerons dans la suite les détails d'implémentation de ce moteur de raisonnement. La quatrième modification est réalisée au niveau de la classe *Comparison* (voir Figure 4.5). C'est au niveau de cette classe que le processus présenté dans la Figure 4.7 est exécuté. Nous avons donc remplacé, pour les comparaisons de catégorie spatiale, le calcul classique de la valeur de similarité par un calcul adaptatif. On effectue d'abord un raisonnement sur les caractéristiques des géométries qui nous fournit les différents

paramètres que nous employons ensuite dans le calcul de la valeur de similarité conformément à l'Équation 4.5.

Les paramètres $\theta, \alpha, \beta, \omega$ et λ sont définis pour chaque paire de géométries à comparer grâce à l'ensemble de règles de décision prédéfinies appliquées aux métadonnées sur les caractéristiques des géométries concernées. Nous avons utilisé le langage RuleML[104] pour la représentation de la base des règles. RuleML est un système unifiant une famille de langages de représentation des règles sur le Web. RuleML fournit une manière simple pour formaliser les règles qui associent des causes et leurs effets. Une règle est composée d'un « corps » qui représente les causes et d'une « tête » qui représente la conséquence. On peut éventuellement déclarer un « fait » en rajoutant une règle sans tête. Un « objectif » est représenté par une règle sans corps. Raisonner sur une base de règles RuleML grâce à un moteur de raisonnement peut être effectué selon deux modes différents : direct (Bottom-Up) ou inverse (Top-Down). Le raisonnement direct permet d'inférer toutes les connaissances possibles à partir de la base de règles. Il peut donc découvrir de nouvelles règles ou de nouveaux faits en composant ceux qui existent déjà dans la base de règles. Un raisonnement inverse est appliqué pour fournir des réponses aux requêtes exprimées par les « objectifs » présents dans la base de règles. Afin de définir les paramètres de comparaison des géométries, nous définissons en amont une base de règles qui permet de les calculer. Un « objectif » est rajouté pour chaque paramètre pour exiger que le raisonnement retourne une valeur pour ce paramètre en sortie. À chaque comparaison d'une paire de géométries, les caractéristiques des deux géométries sont formatées et rajoutées à la base des règles en tant que « faits ». Exécuter un raisonnement inverse (Top-Down) sur l'ensemble des règles de décision, des « faits » et des « objectifs » permet de retourner en sortie les valeurs des différents paramètres de comparaison géométrique. Pour ce faire, nous proposons d'utiliser les moteurs de raisonnement existants.

Cette opération de raisonnement pour le calcul de paramètres est effectuée à chaque comparaison d'une paire de géométries. Une fois les paramètres calculés, la mesure de distance choisie est appliquée entre les deux références spatiales concernées, puis la valeur de similarité est calculée selon l'équation 4.5 en utilisant le seuil θ et les paramètres de convexité et de concavité α et β inférés. Si le paramètre de neutralité λ est retourné avec la valeur « faux », la valeur de similarité calculée représente la similarité effective entre les deux ressources en question dans le cas d'une comparaison monocritère. Dans le cas d'une comparaison multicritère, elle est utilisée ensemble avec le poids ω dans le calcul de la similarité agrégée. Une valeur « vrai » pour le paramètre λ, signifie que la comparaison géométrique est neutralisée, c.-à-d. la similarité géométrique n'est pas prise en compte dans le calcul de la similarité globale. La création d'un lien est impossible entre les deux ressources en question dans le cas d'une comparaison monocritère.

Pour l'implémentation d'un moteur de raisonnement sur les règles de décision, nous nous sommes appuyés sur des outils existants. Nous avons testé deux outils différents : OO jDREW[105] et PROVA[106]. Tout comme PROVA, OO jDREW fournit les deux modes d'exécution : directe et inverse. L'architecture du moteur de raisonnement OO jDREW est beaucoup plus adaptée à une exécution unitaire. Les premiers tests de notre approche avec ce moteur de raisonnement n'étaient donc pas

[104] Rule Markup Language, http://wiki.ruleml.org/
[105] http://www.jdrew.org/oojdrew/
[106] https://prova.ws/

vraiment prometteurs. En effet, lancer plusieurs instances du moteur en parallèle s'est avéré très coûteux en temps et surtout en mémoire consommée. Notre choix s'est arrêté sur PROVA, car les premiers tests de notre implémentation avec ce moteur de raisonnement se sont avérés moins consommateurs en temps de calcul et en espace mémoire. Le problème principal auquel nous avons été confrontés dans l'intégration de PROVA dans Silk est de savoir à quel niveau on doit instancier le moteur de raisonnement, en sachant que plusieurs tâches de mise en correspondance peuvent être exécutées en parallèle. Nous avons donc commencé par la méthode naïve, c.-à-d. en créant une instance du moteur au niveau de chaque comparaison. Bien que cette méthode soit opérationnelle, elle s'avère très inefficace en temps et en mémoire. Nous avons donc pris en compte l'architecture de parallélisation des tâches de mise en correspondance proposée par Silk. Comme nous l'avons expliqué avant, Silk permet de lancer plusieurs *threads* qui constituent des tâches de mise en correspondance qui opèrent chacune sur un bloc différent. L'indépendance des blocs grâce aux méthodes d'indexation garantit l'absence de conflit d'accès aux descriptions des ressources par les tâches de mise en correspondance. Nous avons donc implémenté l'instanciation du moteur de raisonnement au début de chaque tâche de mise en correspondance. Chaque tâche de mise en correspondance utilise une seule instance du moteur de raisonnement pour toutes les comparaisons qu'elle effectue. Il suffit de réinitialiser la base des règles au niveau de chaque comparaison géométrique, en y introduisant les « faits » correspondant aux caractéristiques des références spatiales à comparer. Cette méthode d'implémentation nous a permis de réduire considérablement le temps nécessaire pour exécuter une tâche d'interconnexion. Enfin, pour respecter la généricité de Silk et faciliter l'établissement de la base de règles par l'utilisateur, nous proposons de l'expliciter au sein d'un fichier situé dans le répertoire principal de Silk et modifiable à tout moment indépendamment du code source.

Au-delà des quatre modifications réalisées pour intégrer l'implémentation présentée dans la Figure 4.7, nous avons rajouté quelques autres améliorations afin de garantir le bon déroulement de notre approche. En plus des mesures de distance géométrique déjà implémentées dans Silk (distance entre centroïdes, distance orthodromique, etc.), nous avons rajouté une mesure qui calcule la distance minimale entre un point et n'importe quelle autre primitive géométrique (point, polyligne ou polygone). Cette mesure de distance (voir Figure 4.8) prend en entrée deux géométries exprimées en WKT dans un même système de coordonnées projeté. L'utilisateur peut optionnellement spécifier l'identifiant du système de coordonnées dans le registre EPSG[107] (ex. EPSG:2154 pour spécifier qu'il s'agit de coordonnées en RGF93 / Lambert-93), ou encore, spécifier les coordonnées du rectangle qui englobe toutes les références spatiales des deux sources. Nous avons rajouté une fonction d'indexation qui s'appuie sur l'une de ces deux informations. En effet, les coordonnées du rectangle qui englobe les données ou les coordonnées maximales et minimales du système de coordonnées renseigné sont utilisées pour fixer les valeurs maximales et minimales des coordonnées à indexer. Une valeur d'indexation est ensuite calculée pour chacun des deux axes en utilisant la méthode d'indexation des valeurs numériques expliquée précédemment sur chacune des coordonnées du centroïde de la géométrie concernée. Grâce à l'opération de conjonction d'index proposée par Silk, les deux valeurs sont fusionnées pour affecter la géométrie à un seul bloc final. Même si la valeur du seuil utilisée par la fonction de similarité est calculée dynamiquement, une valeur de seuil primaire doit être renseignée pour être utilisée dans le calcul de l'index.

[107] http://www.epsg.org/

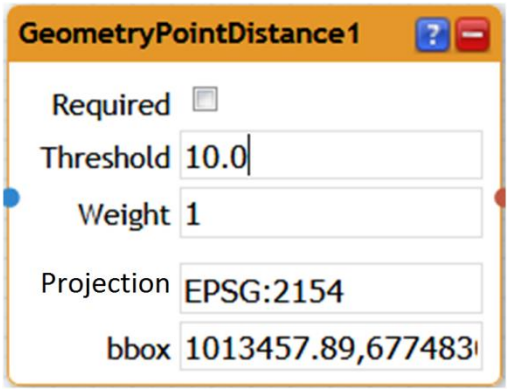


Figure 4.8 Exemple de l'usage de la mesure de distance géométrique rajoutée à Silk. Le code du système de coordonnées de référence ainsi que les coordonnées maximum et minimum du rectangle englobant sont fournis.

Afin de garantir que la mesure de distance soit appliquée entre des géométries représentées dans un même système de coordonnées projeté, nous avons également mis en œuvre un opérateur de transformation de systèmes de coordonnées. Il suffit de renseigner les codes EPSG des systèmes de coordonnées d'origine des données et celui vers lequel on veut les projeter pour que l'opérateur effectue la transformation (ex. voir Figure 4.9). Cette transformation est réalisée une seule fois pour chaque géométrie lors du chargement des données.

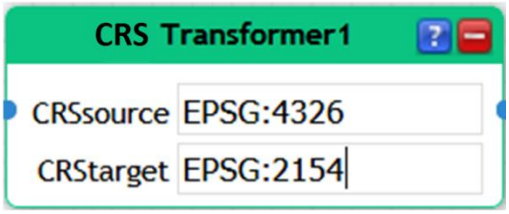


Figure 4.9 Exemple de l'utilisation de l'opérateur de transformation de coordonnées géographiques. Dans cet exemple, les coordonnées seront transformées du WGS84 vers Lambert93

Plusieurs tests ont été réalisés pendant l'implémentation de cette approche pour vérifier son bon déroulement. Nous présentons dans la suite un cas d'application réel où on cherche à interconnecter des données géoréférencées par des références spatiales dont les caractéristiques sont hétérogènes à la fois entre les différentes sources de données, mais également au sein de ces sources de données.

4.1.4 Test et validation : interconnexion de monuments de la ville de Paris

Afin d'évaluer notre approche d'adaptation dynamique de paramètres de comparaison, nous avons choisi deux sources de données qui décrivent les monuments historiques de Paris. La première source est la base de données Mérimée[108]. Cette base, produite et maintenue par le ministère de la Culture et de la Communication, recense les édifices protégés au titre des monuments historiques en France. Dans cette base les monuments sont géoréférencés par des références spatiales indirectes : chaque monument est décrit par deux attributs textuels qui représentent son adresse ainsi que le code INSEE de la commune dans laquelle il se trouve. Chaque monument peut avoir plusieurs

[108] https://www.data.gouv.fr/fr/datasets/liste-des-immeubles-proteges-au-titre-des-monuments-historiques/

adresses. Étant fournie sous forme de fichier CSV, nous avons transformé cette base en données RDF en utilisant la plateforme Datalift. La base de données Mérimée couvre les monuments historiques de tout le territoire français, nous avons sélectionné seulement ceux se trouvant dans Paris. Nous avons ensuite géocodé les adresses des monuments sélectionnés en utilisant la base de données BD ADRESSE®. Ceci est réalisé par interconnexion entre les ressources de la source Mérimée et celles de la BD ADRESSE®, également transformée en RDF en utilisant la plateforme Datalift (cf. section 3.2.1). Pour cette interconnexion, nous avons utilisé trois comparaisons. La première comparaison est employée entre les noms de voies dans les deux sources. Elle utilise une mesure de distance de de type « sac de mots », fondée sur une mesure de *Levenshtein* et qui ignore l'ordre des mots[109], car l'ordre des mots n'est pas le même dans les deux noms de voies comparées. Nous avons fixé le seuil de la distance de *Levenshtein* à 1 car au-delà de cette valeur la mise en correspondance n'est plus précise. La seconde et la troisième comparaison cherchent une égalité exacte pour les valeurs des numéros de rue et des codes INSEE entre les deux sources. Les trois comparaisons sont agrégées par une moyenne arithmétique. Seuls les liens ayant un score global supérieur à 0,7 ont été concervés. Cette interconnexion nous a permis d'obtenir 1582 monuments géoréférencés. Les coordonnées des géométries des monuments Mérimée, récupérées de la BD ADRESSE®, sont exprimées dans le système de coordonnées de référence Lambert93. Nous avons évalué les résultats du géocodage en vérifiant la précision et le rappel d'une partie des données sur cet echantillon : le géocodage a une précision de 100% et un rappel de 98%. La deuxième source de données rassemble 625 ressources du chapitre français de DBpedia qui représentent les monuments de Paris. Il s'agit des données que nous avons traitées précédemment dans la partie 3.2.2. Dans ce jeu de données, les coordonnées des ressources sont exprimées dans le système de coordonnées de référence WGS84.

Préparation des données

Comme nous l'avons vu dans la section 3.2, les références spatiales des monuments Mérimée (récupérées à partir de la BD ADRESSE®) tout comme celles des monuments DBpedia présentent des hétérogénéités internes à chacune des sources, et par conséquent, des hétérogénéités entre les deux sources.

Nous avons réussi, dans la section 3.2.1, à extraire à partir des métadonnées disponibles, la modélisation géométrique ainsi que la précision planimétrique des références spatiales des ressources de la BD ADRESSE®. Les références spatiales des adresses ainsi que leurs caractéristiques exprimées dans le vocabulaire de la sémantique des XY sont importées et intégrées dans le jeu de données Mérimée : chaque ressource Mérimée est désormais liée à une géométrie issue de la BD ADRESSE®, elle-même liée à la description de ses caractéristiques. Elle peut être liée à plusieurs points dans le cas où le monument est géoréférencé par plusieurs adresses.

Nous avons également montré dans la section 3.2.2 comment nous avons extrait les caractéristiques des géométries de chaque ressource représentant les monuments historiques dans DBpedia en utilisant une approche de classification par apprentissage supervisé qui s'appuie sur des données topographiques de référence.

[109] *tokenwiseDistance* dans https://github.com/silk-framework/silk/blob/master/doc/Plugins.md#tokenbased

Ainsi, nous disposons de deux jeux de données représentant le même type d'entités géographiques du monde réel, et dont les ressources sont géoréférencées par des géométries décrites par l'ensemble de leurs caractéristiques selon le vocabulaire de la sémantique des XY. Les jeux de données réunissent les conditions nécessaires pour tester l'implémentation de notre approche.

Tests d'évaluation comparatifs

Afin d'évaluer notre approche, nous devons comparer ses résultats avec l'approche classique non adaptative qui opère selon une configuration fixée avant l'interconnexion. Pour cela nous utilisons la version de Silk qui implémente notre approche et nous la comparons avec la version originale de Silk. L'amélioration que nous cherchons principalement à apporter porte sur la qualité des résultats d'interconnexion par rapport à l'approche classique. Les résultats d'interconnexion sont ensuite comparés à l'aide des mesures de précision et de rappel et surtout de F-mesure qui constitue le compromis entre la précision et le rappel. Nous verrons donc dans la suite comment nous avons constitué un jeu de relations de correspondance de référence pour nos jeux de données de test, puis nous présenterons les tests effectués avec une approche d'interconnexion dotée d'un paramétrage initial fixe et avec notre approche à paramétrage adaptatif. Enfin, nous discuterons des résultats obtenus.

Création d'un mapping de référence

Disposer d'un *mapping de référence*[110] est nécessaire afin de pouvoir évaluer les résultats d'interconnexion. Nous commencons donc par la création de cet ensemble de liens de référence entre monuments DBpedia et monuments Mérimée. Bien que les pages décrivant les monuments historiques français dans Wikipedia précisent que ceux-ci font partie du classement Mérimée, les ressources DBpedia décrivant ces monuments historiques ne disposent pas dans leurs descriptions de références directes vers les objets Mérimée correspondants. Nous avons constaté cependant que la source de données Wikidata[111] fournit des références entre ressources issues de Wikipedia et d'autres bases du Web (voir Figure 4.10). Ceci concerne la plupart des ressources qui décrivent des monuments historiques français, pour lesquelles les identifiants des enregistrements correspondants dans Mérimée sont fournis. Nous avons donc choisi de nous appuyer sur ces informations pour constituer notre jeu de relations de correspondance de référence.

Grâce à l'API[112] de Wikidata, nous avons récupéré les identifiants Mérimée des ressources DBpedia en utilisant la liste des labels DBpedia comme moyen de recherche dans Wikidata. Nous avons parcouru les résultats obtenus en les traduisant en liens entre les ressources DBpedia et les ressources Mérimée. Nous avons ainsi créé partiellement un jeu de relations de correspondance de référence. Nous avons ensuite créé une application de visualisation cartographique de ces liens de référence (voir Figure 4.11) afin de pouvoir les vérifier et les compléter en analysant les descriptions textuelles des ressources dans les deux sources. L'architecture de cette application est semblable à celle que nous présentons dans la section 4.3.2. Nous avons finalement réussi à récupérer 445 liens de référence que nous utiliserons dans la suite pour l'évaluation des résultats d'interconnexion.

[110] jeu de relations de correspondance de référence
[111] https://www.wikidata.org/
[112] https://www.mediawiki.org/wiki/API

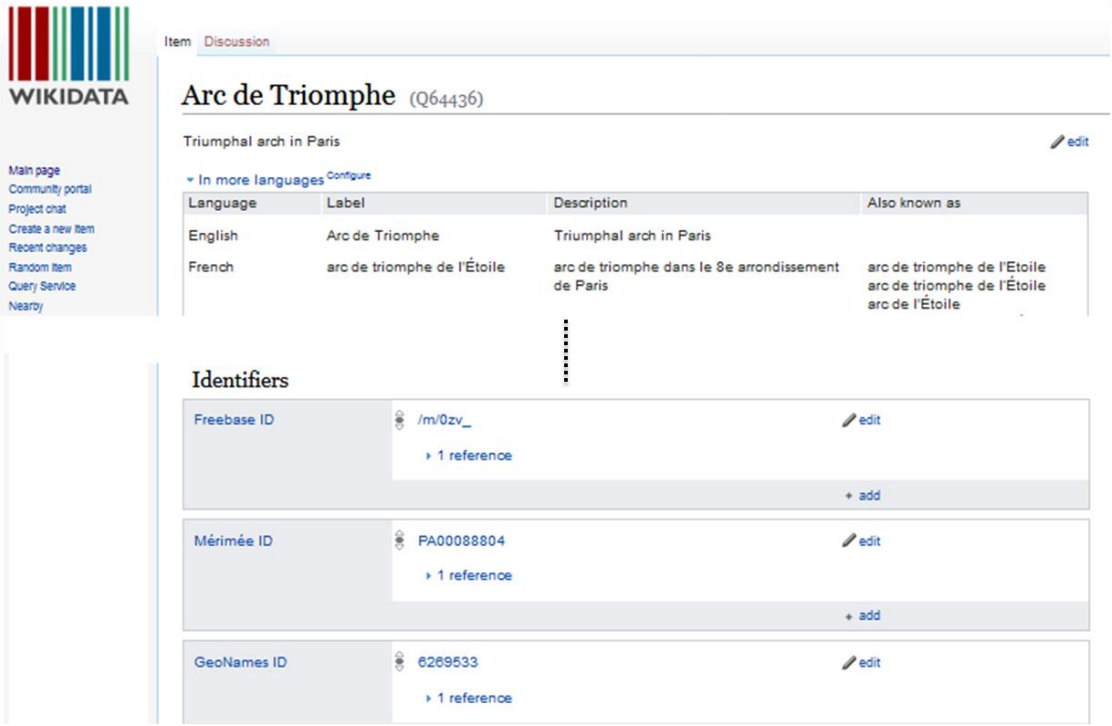


Figure 4.10 Exemple d'une ressource Wikidata décrivant un monument historique parisien (l'Arc de Triomphe). Sa description comprend une référence vers l'identifiant du monument dans la base Mérimée.

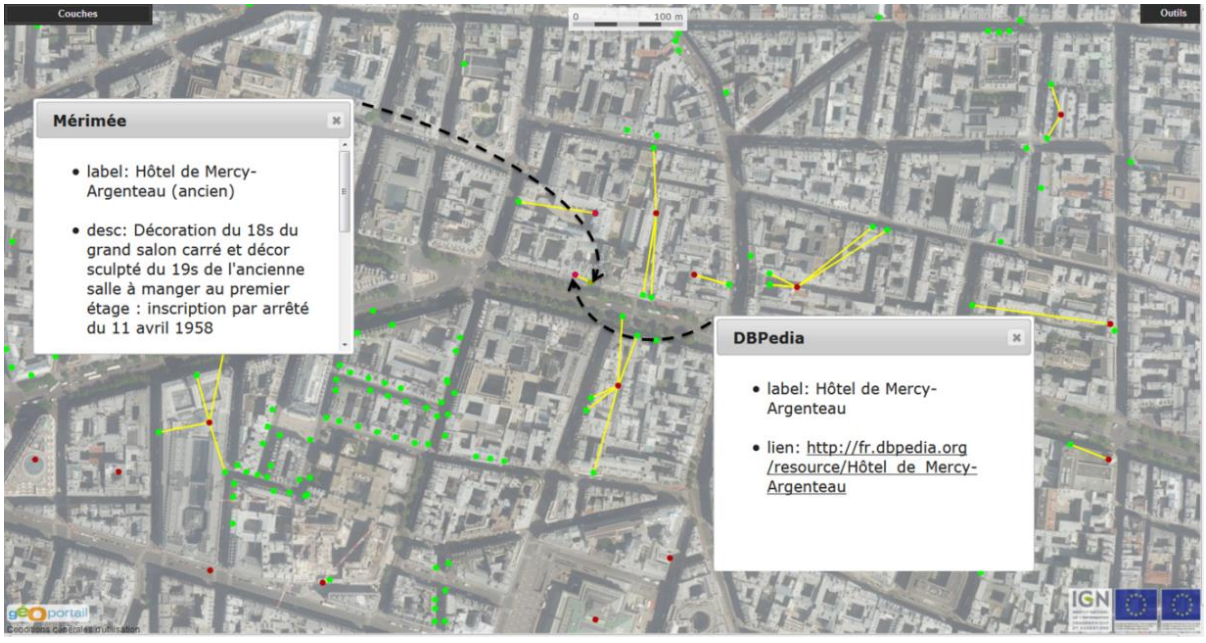


Figure 4.11 Interface de visualisation de l'ensemble de liens de référence. Les points rouges représentent les monuments DBpedia, les verts représentent les monuments Mérimée. Les lignes jaunes représentent les liens. En cliquant sur un point, une fenêtre récupère à la volée et affiche quelques informations sur le monument concerné.

Dans ce cas d'application, nous nous sommes concentrés sur l'utilisation de la comparaison géométrique comme seul critère d'interconnexion. En effet, les seules autres propriétés qui nous semblaient comparables sont les noms des monuments, mais ceux-ci ne sont pas fiables, car une grande partie des monuments Mérimée ont des noms génériques (ex. « Immeuble ») alors que les

labels des monuments DBpedia ont des noms spécifiques. Ainsi, en utilisant seulement la comparaison géométrique, nous avons testé notre approche ainsi que l'approche classique avec laquelle nous comparons nos résultats.

Mise en œuvre d'une approche d'interconnexion classique de Silk

Pour ce premier test d'interconnexion, nous avons utilisé la version originale de Silk 2.6. Celle-ci calcule la valeur de similarité pour chaque comparaison selon l'équation 4.4. Cette version n'implémente pas notre approche d'adaptation dynamique des paramètres de comparaison géométrique. Les ressources des monuments étant géoréférencées par des points dans les deux jeux de données, nous avons préparé un script de configuration d'interconnexion en précisant que nous utilisons une distance euclidienne comme mesure de comparaison. Pour cela, nous avons employé la mesure de distance que nous avons rajoutée à Silk (cf. Figure 4.8). Nous avons également appliqué l'opérateur de transformation de coordonnées que nous avons rajouté dans Silk (cf. Figure 4.9) afin de projeter les coordonnées des monuments DBpedia dans la même projection que les coordonnées des monuments Mérimée, c.-à-d. le Lambert93.

Le seul paramètre sur lequel nous pouvons intervenir lors de la configuration de cette interconnexion est le choix du seuil de distance. Nous avons donc varié la valeur du seuil en évaluant pour chaque valeur de seuil la qualité des liens. Ceci est réalisé dans le but de trouver la valeur du seuil qui fournit le meilleur compromis entre précision et rappel, c.-à-d. la meilleure F-mesure. Les résultats d'évaluation sont présentés dans le Tableau 4.1 et la Figure 4.12.

En augmentant le seuil, le rappel augmente alors que la précision diminue. Le meilleur compromis entre les deux valeurs correspond à une valeur de F-mesure égale à **58,42%** obtenue pour seuil égal à 40m. C'est avec cette valeur de F-mesure que nous comparerons les résultats obtenus par notre approche.

Seuil	Précision	Rappel	F-mesure
10	84,55%	20,90%	33,51%
20	74,15%	39,33%	51,40%
30	64,15%	51,46%	57,11%
40	57,96%	58,88%	**58,42%**
50	50,09%	63,15%	55,86%
60	44,75%	65,17%	53,06%
70	40,43%	66,97%	50,42%
80	36,76%	68,31%	47,80%
90	33,99%	69,44%	45,64%
100	31,68%	70,34%	43,68%

Tableau 4.1 Évaluation des résultats de l'approche classique (sans adaptation dynamique de paramètres) en termes de précision, de rappel et de F-mesure selon les valeurs de seuil.

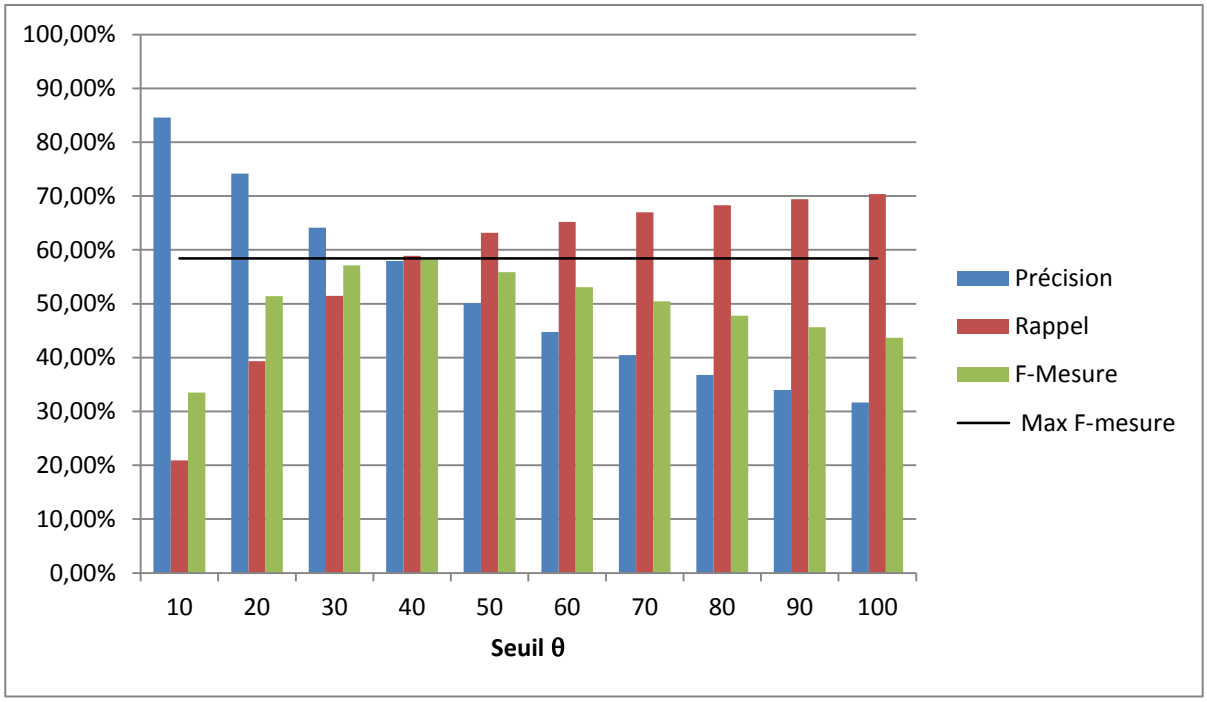


Figure 4.12 Résultats de l'approche classique (sans adaptation dynamique de paramètres) en termes de précision, de rappel et de F-mesure selon les valeurs de seuil. La valeur maximale de la F-mesure sera utilisée comme repère pour les résultats notre approche.

Mise en œuvre de l'approche d'adaptation dynamique des paramètres de comparaison

Nous avons également appliqué notre approche d'interconnexion à paramétrage adaptatif dynamique à ces jeux de données sur les monuments historiques parisiens. Notre approche nécessite la définition d'une base de règles qui permet de générer les valeurs des différents paramètres de comparaison θ, α, β, ω et λ pour chaque paire de géométries dont on cherche à évaluer la similarité. Vu qu'il s'agit d'un cas d'interconnexion monocritère, le paramètre du poids ω ne sera pas utile et sera donc défini comme constant dans la base de règles. Le paramètre β, qui permet de rendre convexe ou concave la partie négative de la fonction de similarité, ne sera pas utile non plus, car les valeurs de similarité négative (quand la distance est supérieure au seuil) n'ont d'utilité que dans le cas d'une interconnexion multicritère. La mesure de distance est également fixée à une mesure de distance euclidienne, car il ne s'agit dans ce cas que de simples points, décrit par des coordonnées planes. A chaque comparaison, les caractéristiques des deux références spatiales sont formatées avant d'être rajoutées à la base de règles pour le raisonnement.

Notre première intuition est d'adapter le calcul du seuil de distance θ, utilisé dans le calcul de la fonction de similarité, selon les valeurs de précision planimétrique et la différence de modélisation géométrique. Ainsi, nous avons défini la base de règles suivante (Encadré 4.5):

thêta (X) ← hôte$_1$ = hôte$_2$, elemCarac$_1$ = elemCarac$_2$, X=préc$_1$ + préc$_2$.
thêta (X) ← hôte$_1$ = hôte$_2$, elemCarac$_1$!= elemCarac$_2$, X= préc$_1$ + préc$_2$+ delta$_e$.
thêta (X) ← hôte$_1$!= hôte$_2$, X= préc$_1$ + préc$_2$ + delta$_h$.

alpha(1). beta(1). oméga(1). lambda(0). distance("euclidian").

← thêta (). ←alpha(). ←beta(). ←oméga(). ←lambda(). ← distance().

Encadré 4.5 Base de règle BR$_1$

Les éléments « elemCarac$_i$ » et « *hôte$_i$* » (avec $i \in \{1,2\}$) représentent respectivement l'élément caractéristique de la forme et l'hôte choisis pour la représentation géométrique de la référence spatiale (cf. section 3.1.2). Les règles écrites en noir permettent de calculer le seuil θ à partir des caractéristiques des références spatiales comparées. La première ligne signifie que le seuil de distance est calculé à partir de la somme des valeurs de précision planimétriques (« *préc$_i$* ») si les deux références spatiales partagent la même modélisation géométrique, c.-à-d. même élément caractéristique de la forme et même hôte. La deuxième règle s'applique quand les deux références spatiales ont des éléments caractéristiques de la forme différents saisis par rapport à un même hôte (ex. point saisi vers le centre d'un bâtiment comparé à un autre point saisi vers la façade du bâtiment). Dans ce cas, le seuil est estimé comme la somme des deux précisions planimétriques des deux références spatiales, en lui rajoutant un biais « *delta$_e$* » pour compenser la différence d'élément caractéristique de la forme. La troisième règle est appliquée quand les deux références spatiales ont des modélisations géométriques complètement différentes. Dans ce cas un biais « *delta$_h$* » (plus important que « *delta$_e$* ») est rajouté à la somme des précisions planimétriques pour estimer le seuil. La ligne verte rassemble des « faits » qui permettent de fixer les paramètres autres que le seuil comme expliqué dessus. La ligne rouge rassemble l'ensemble des « objectifs » qui représentent les paramètres à renvoyer en sortie du raisonnement. Ainsi, nous avons établi la première base de règles BR$_1$ en fixant *delta$_e$* et *delta$_h$* respectivement à 10m et 15m. Le résultat d'interconnexion utilisant cette base de règles est présenté dans le Tableau 4.2 et la Figure 4.13.

En nous appuyant sur l'expérience des résultats obtenus avec la base de règle BR$_1$, nous avons défini une deuxième base de règles en rajoutant d'autres règles plus fines qui permettent d'adapter le paramètre α de convexité de la fonction de similarité ainsi que le paramètre de neutralisation λ. Nous avons également augmenté les valeurs *delta$_e$* et *delta$_h$* à 20m et 30m.

thêta (X) ← hôte$_1$ = hôte$_2$, elemCarac$_1$ = elemCarac$_2$, X=préc$_1$ + préc$_2$.
thêta (X) ← hôte$_1$ = hôte$_2$, elemCarac$_1$!= elemCarac$_2$, X= préc$_1$ + préc$_2$+ delta$_e$.
thêta (X) ← hôte$_1$!= hôte$_2$, X= préc$_1$ + préc$_2$ + delta$_h$.

alpha (1) ← hôte$_1$ = hôte$_2$, elemCarac$_1$ = elemCarac$_2$.
alpha (2) ← hôte$_1$ = hôte$_2$, elemCarac$_1$!= elemCarac$_2$.
alpha (3) ← hôte$_1$!= hôte$_2$.

lambda (1) ← hôte$_1$ = « zone d'adressage ».
lambda (1) ← hôte$_2$ = « zone d'adressage ».
lambda (1) ← hôte$_1$ = « commune ».
lambda (1) ← hôte$_2$ = « commune ».
lambda (0) ← hôte$_1$!= « zone d'adressage », hôte$_2$!= « zone d'adressage », hôte$_1$!= « commune », hôte$_2$!= « commune ».

beta(1). oméga(1). distance(''euclidian'').

← thêta (). ←alpha(). ←beta(). ←oméga(). ←lambda(). ← distance().

Encadré 4.6 Based de règles RB$_2$

Les règles rajoutées permettent d'augmenter la convexité de la fonction de similarité quand les références spatiales ont des modélisations géométriques différentes. Ceci veut dire que la fonction de similarité est beaucoup plus stricte quand il s'agit de deux références spatiales ayant la même modélisation géométrique. La valeur de similarité est complètement neutralisée quand la modélisation géométrique d'une des deux références spatiales comparées est établie par rapport à un objet géographique hôte très vague ou très large (commele centre de la commune ou le centre d'une zone d'adressage). Ceci suppose que le point n'est pas situé à proximité de son édifice correspondant et l'utiliser comme critère de comparaison ne serait pas fiable pour établir un lien. Ainsi nous avons établi la base de règles BR_2 suivante (règles rajoutées en gris), présentée dans l'Encadré 4.6.

Les lignes verte et rouge représentent toujours les paramètres fixés et les requêtes à résoudre. Les résultats d'interconnexion en utilisant BR_2 sont également présentés dans le Tableau 4.2 et la Figure 4.13.

Base de règle	Précision	Rappel	F-mesure
BR_1	70,43%	58,88%	64,14%
BR_2	71,43%	62,92%	**66,91%**

Tableau 4.2 Évaluation des résultats d'interconnexion par l'approche d'adaptation dynamique des paramètres de comparaison géométrique.

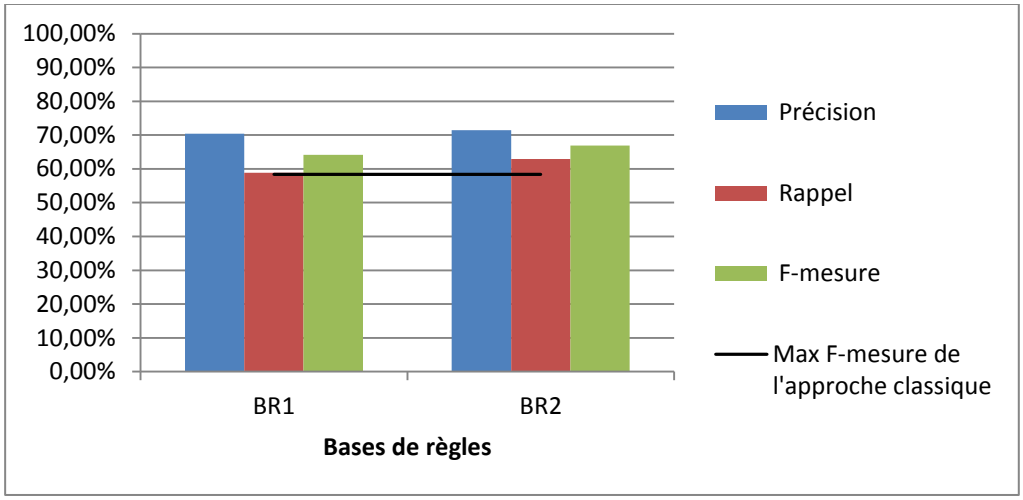


Figure 4.13 Résultats d'interconnexion par l'approche d'adaptation dynamique des paramètres de comparaison géométrique.

Ces résultats montrent que notre approche génère des résultats dont les valeurs de F-mesure sont meilleures que celles obtenues par l'approche classique. Ces résultats ont été obtenus avec un jeu de données DBpedia dont les caractéristiques des références spatiales sont obtenues en appliquant la méthode d'acquisition par apprentissage supervisé présentée dans la section 3.2.2. Plus précisément, les caractéristiques des géométries correspondent aux résultats obtenus par l'utilisation des réseaux bayésiens (cf. première ligne du tableau 3.2). Afin d'étudier l'effet de l'erreur engendrée par l'utilisation de la méthode d'acquisition des caractéristiques des géométries par apprentissage sur les résultats d'interconnexion, nous avons testé notre approche après avoir corrigé les erreurs d'apprentissage et réévalué les valeurs de précision planimétriques, c.-à-d. dans un cas parfait où l'apprentissage obtient une précision et un rappel de 100%. Les résultats obtenus dans ce cas de figure pour les bases de règles BR_1 et BR_2 sont présentés dans le Tableau 4.3 et la Figure 4.14.

Base de règle	Précision	Rappel	F-mesure
BR_1	73,46%	59,10%	65,50%
BR_2	70,43%	62,70%	**66,59%**

Tableau 4.3 Évaluation des résultats d'interconnexion par l'approche d'adaptation dynamique des paramètres de comparaison géométrique après correction des résultats d'acquisition automatique des caractéristiques des géométries.

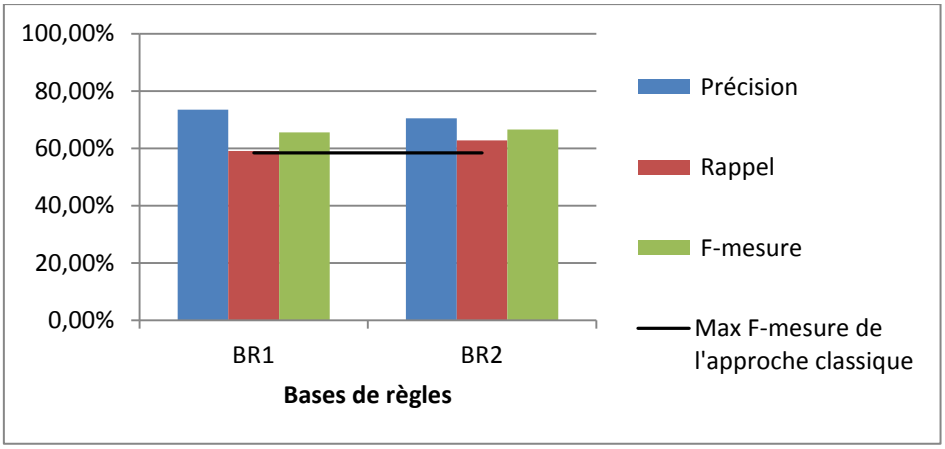


Figure 4.14 Résultats d'interconnexion par l'approche d'adaptation dynamique des paramètres de comparaison géométrique après correction des résultats d'acquisition automatique des caractéristiques des géométries.

Discussion des résultats d'interconnexion

Les résultats obtenus en faisant varier un seuil de distance fixe dans une approche d'interconnexion classique sont tout à fait logiques. Les valeurs de seuil les plus basses assurent une précision élevée, car elles ne permettent d'interconnecter que des ressources qui sont spatialement très proches et qui sont les plus susceptibles d'être équivalentes : elles réduisent le nombre de liens « faux positifs ». Cependant, ces valeurs basses ne favorisent pas un rappel important, car elles réduisent également le nombre de liens « vrais positifs ». Accroître la valeur du seuil permet de découvrir plus de liens, ce qui permet d'augmenter le nombre de liens vrais positifs et donc d'augmenter le rappel, au détriment de la précision qui diminue à cause du nombre de liens faux positifs qui augmente également. La décision de favoriser une précision optimale des liens résultants plutôt qu'un rappel optimal, ou l'inverse, dépond des l'objectif de l'interconnexion. De manière générale, l'objectif de l'interconnexion est de trouver le meilleur compromis entre ces deux valeurs, *c.-à-d.* découvrir le maximum de liens vrais positifs, tout en ayant un minimum de liens faux positifs. Optimiser la valeur de la F-mesure, qui constitue une moyenne harmonique entre les valeurs de précision et de rappel, permet de trouver le meilleur compromis entre ces deux dernières. Comme nous l'avons expliqué dans les résultats d'interconnexion par l'approche classique, la valeur optimale de la F-mesure est obtenue pour nos jeux de données en utilisant un seuil de distance de valeur 40m. C'est avec cette valeur de F-mesure que nous avons comparé les résultats de notre approche.

Les résultats obtenus par notre approche adaptative montrent une amélioration claire de la valeur de la F-mesure. Adapter la valeur du seuil de la fonction de similarité aux différences des caractéristiques des références spatiales permet de profiter des avantages des petites valeurs ainsi que de ceux des grandes valeurs de seuil. Notamment, en utilisant la base de règles BR_2 nous avons réussi à augmenter le nombre de liens vrais positifs de 6% tout en évitant 40% des liens faux positifs du meilleur résultat de l'approche classique. La Figure 4.15 montre un exemple de lien qu'on a réussi

à découvrir grâce à l'adaptation de la valeur du seuil. Les deux points géoréférencent le même monument du monde réel « l'Hôtel de la Marine » dans DBpedia[113] et dans Mérimée[114]. Comme le montre la figure, la distance entre les deux points est de 50m : ce lien n'a pas pu être découvert dans le cas optimal de l'approche classique en raison de seuil de distance maximale de 40m. Cependant dans l'approche adaptative le seuil a été calculé dynamiquement en prenant en compte les valeurs des précisions planimétriques des deux points (~23m pour le monument DBpedia et 12m pour le monument Mérimée) ainsi que la différence de modalisations géométriques (la géométrie du monument DBpedia est modélisée au barycentre du bâtiment alors que celle du monument Mérimée est modélisée par un point de la façade du bâtiment). Dans ce cas, dans la base de règles RB_2, c'est la règle suivante qui a été appliquée pour calculer la valeur du seuil (pour $delta_e$ = 20m) :

thêta (X) ← hôte₁ = hôte₂, elemCarac₁ != elemCarac₂, X= préc₁ + préc₂+ deltaₑ .

La valeur du seuil de distance dans ce cas est donc de 55m ce qui permet de découvrir un lien entre ces deux ressources.

La Figure 4.16 présente un exemple de lien erroné découvert par l'approche classique, mais qu'on a réussi à éviter avec l'approche adaptative. Il s'agit dans ce cas de deux points qui localisent deux monuments complètement distincts : le monument DBpedia (point rouge) représente l'immeuble de « la sous-station Opéra »[115] alors que le monument Mérimée représente un autre immeuble classé[116]. Comme le montre la figure, la distance entre les deux points est à peu près de 37m. Cette valeur étant inférieure à 40m, un lien est donc créé entre ces deux monuments par l'approche classique. Avec l'approche adaptative, le seuil de distance est calculé dynamiquement: il s'agit ici de deux points ayant la même modélisation géométrique (sur la façade du bâtiment) et des précisions planimétriques d'environ 15m pour le point du monument DBpedia et de 12m pour le point du monument Mérimée. Vu que les deux points partagent la même modélisation géométrique, ils ne sont pas censés être très loin s'ils localisent le même monument. C'est la règle suivante de la base de règles BR_2 qui s'applique pour calculer le seuil :

thêta (X) ← hôte₁ = hôte₂, elemCarac₁ != elemCarac₂, X= préc₁ + préc₂ .

Le seuil de distance prend donc dans ce cas une valeur de 27m, ce qui permet d'éviter la création d'un lien erroné.

[113] http://fr.dbpedia.org/resource/Hôtel_de_la_Marine
[114] Référence Mérimée: « PA00088817 »
[115] http://fr.dbpedia.org/page/Sous-station_Opéra
[116] Référence Mérimée : « PA00088933 »

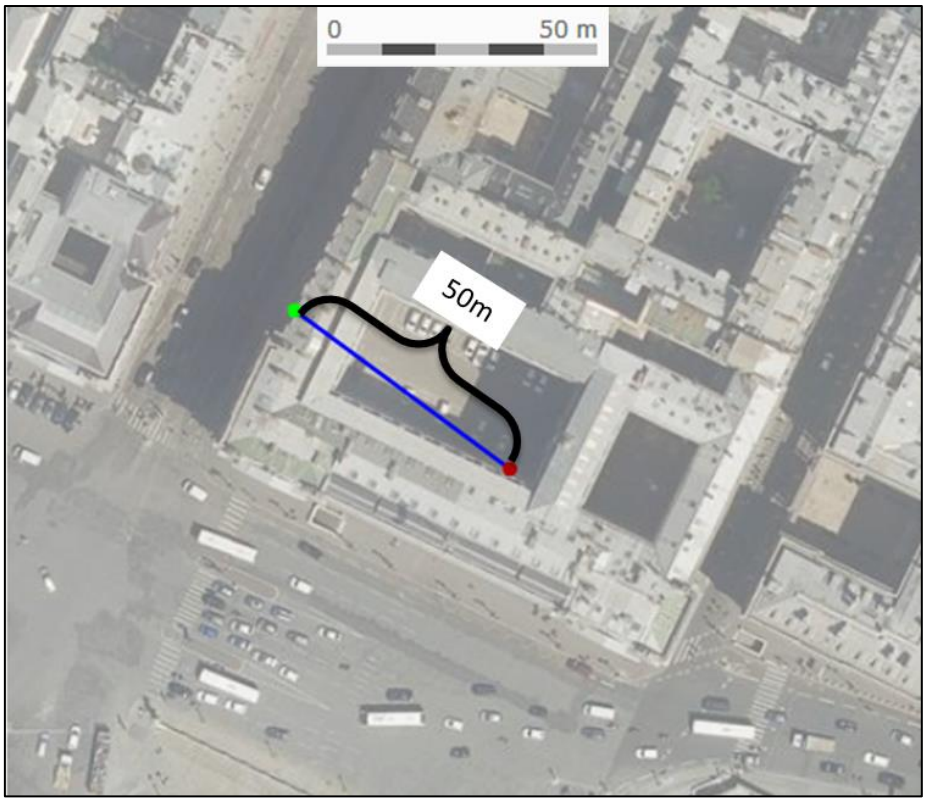


Figure 4.15 Lien découvert seulement par l'approche adaptative. Les deux monuments représentent « l'Hôtel de la Marine » dans DBpedia (point rouge) et Mérimée (point vert).

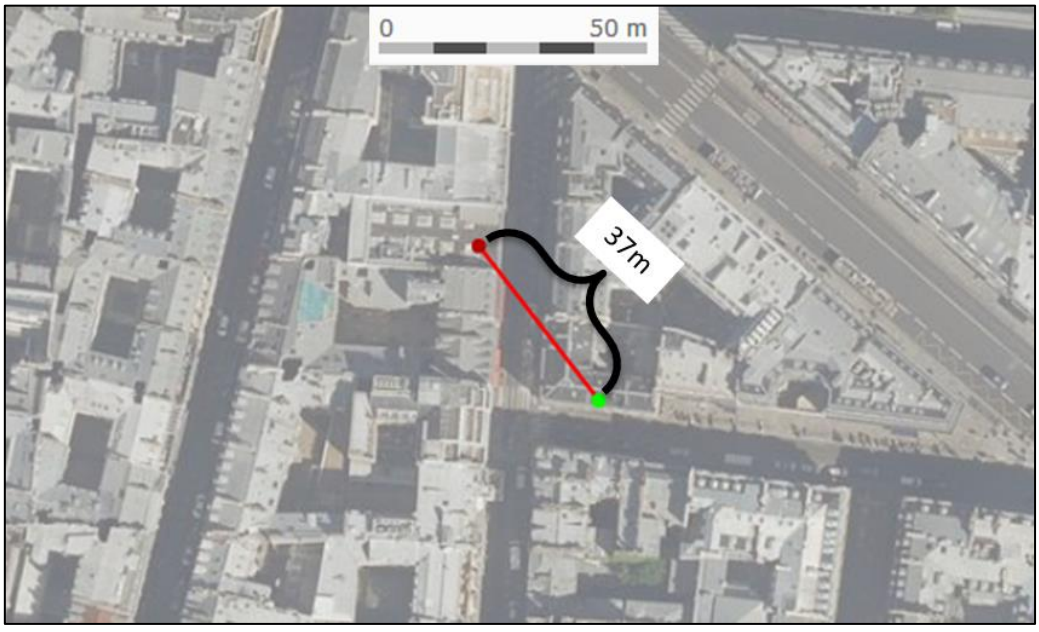


Figure 4.16 Faux lien (entre deux monuments distincts) découvert par l'approche classique, mais évité par l'approche adaptative. Le point rouge représente le monument DBpedia. Le point vert représente le monument Mérimée.

La comparaison des résultats présentés dans le Tableau 4.3 par rapport à ceux du Tableau 4.2 montre que les erreurs d'apprentissage du type de modélisation géométrique (*cf.* 3.2.2) engendrent un effet sur le résultat d'interconnexion. Cependant nous constatons que cet effet reste léger même pour les résultats d'apprentissage les moins bonss (tels que présentés dans le tableau 3.2).

En termes de complexité temporelle, il n'est pas surprenant que l'approche adaptative (9~14 secondes) nécessite plus de temps d'exécution que l'approche classique (~2 secondes). En effet, dans sa version naïve, un processus d'interconnexion exécuté par Silk effectue un produit cartésien pour la comparaison de deux sources de données, résultant en une complexité temporelle quadratique. Grâce aux techniques d'indexation et de blocage proposées, Silk réduit la complexité temporelle de l'approche d'interconnexion classique en une complexité quasi-linéaire. L'implémentation du raisonneur dans l'approche adaptative augmente la complexité temporelle. En effet, l'implémentation du moteur de raisonnement PROVA présente une complexité temporelle linéaire sur le nombre de règles dans la base de règles. L'intégrer dans Silk pour être exécuté au niveau de chaque comparaison de géométries engendre une complexité totale plus élevée. Le temps d'exécution demeure quand même raisonnable, car nous avons définit un nombre minimal de règles qui permettent d'adapter d'une manière efficace le paramétrage des comparaisons géométriques. Une base de règles beaucoup plus détaillée et plus spécifique dans la gestion des différences des caractéristiques des géométries pourrait générer de meilleurs résultats d'interconnexion, mais serait certainement moins efficiente en temps d'exécution. Il est cependant nécessaire que les règles soient décidables, *c.-à-d.* qu'elles soient suffisantes pour générer une valeur unique pour chaque paramètre demandé en sortie du raisonnement.

Dans l'exemple des monuments que nous avons présenté, nous avons appliqué notre approche d'adaptation dynamique dans une interconnexion monocritère. Toutefois, elle est toujours employable dans le cadre d'une interconnexion multicritère. Dans cet exemple nous avons cherché uniquement des liens de cardinalité 1:1. Nous avons cependant remarqué que des liens de cardinalité n:m existent entre les monuments des deux jeux de données. Nos premiers tests pour découvrir des liens de cardinalité n:m ont montré une perte considérable en précision. En effet, chercher des relations de cardinalités n:m revient à créer, pour chaque ressource du jeu de données source, des liens avec les ressources du jeu de données cible se trouvant dans un rayon égal au seuil. On ne garde donc pas seulement la ressource la plus proche. Dans un cas d'application à forte densité de points comme celui qui est présenté ici, ceci augmente drastiquement le nombre de liens faux positifs. Nous proposons dans la suite une approche complémentaire qui s'appuie sur les données topographiques de référence pour chercher des liens de cardinalité n:m.

4.2 Utilisation d'un référentiel topographique comme support pour l'interconnexion de données géoréférencées

Afin de pouvoir compléter les résultats obtenus par l'approche proposée dans 4.1, nous proposons une autre approche qui vise à découvrir des liens de cardinalité n:m tout en réduisant l'effet des hétérogénéités géométriques sur l'interconnexion de données géoréférencées. L'idée de l'approche est de s'appuyer sur un jeu de données topographiques de référence comme ressource de support pour l'interconnexion. Nous fournissons en premier lieu dans cette partie une description générale de cette approche. Nous décrivons ensuite la mise en œuvre de cette approche appliquée au cas des monuments historiques de Paris pour la valider.

4.2.1 Description générale de l'approche

Comme nous l'avons vu précédemment, les hétérogénéités géométriques qui peuvent exister entre les données géoréférencées sont liées principalement aux différences de niveaux de détail et de

règles de saisie des géométries. Utiliser la comparaison des géométries comme critère pour l'interconnexion de données dans le Web peut devenir inefficace : par exemple les géométries de deux ressources qui représentent la même entité du monde réel peuvent être très éloignées, ou les géométries de deux ressources qui représentent des entités distinctes du monde réel peuvent être très proches. Nous avons montré dans le chapitre 3 comment on peut utiliser des données géographiques de référence pour identifier les différentes caractéristiques de chaque géométrie dans un jeu de données. Ces caractéristiques sont ensuite utilisées dans l'adaptation dynamique des paramètres d'interconnexion telle que présentée dans l'approche précédente (cf. section 4.1). Nous avons appliqué l'approche adaptative précédente dans un contexte de recherche de liens de cardinalité **1:1**. Pour trouver des liens de cardinalité **n:m**, nous proposons dans cette approche complémentaire de nous appuyer sur les relations entre les ressources à interconnecter et les données topographiques de support qui décrivent leur contexte géographique. Similairement à l'approche proposée par (Aleksovski et al., 2006) dans le domaine de l'alignement d'ontologies, nous suggérons d'**ancrer** les ressources à interconnecter à leurs entités géographiques correspondantes représentées au sein d'un même jeu de données de référence, puis **dériver** des liens d'équivalence (ou autres) entre ces ressources à partir des relations d'ancrage. La Figure 4.17 représente un exemple de dérivation de relation d'équivalence entre deux ressources en utilisant les liens créés entre elles et leurs entités géographiques correspondantes via une approche d'interconnexion fondée sur l'utilisation de références spatiales comme critère de mise en correspondance.

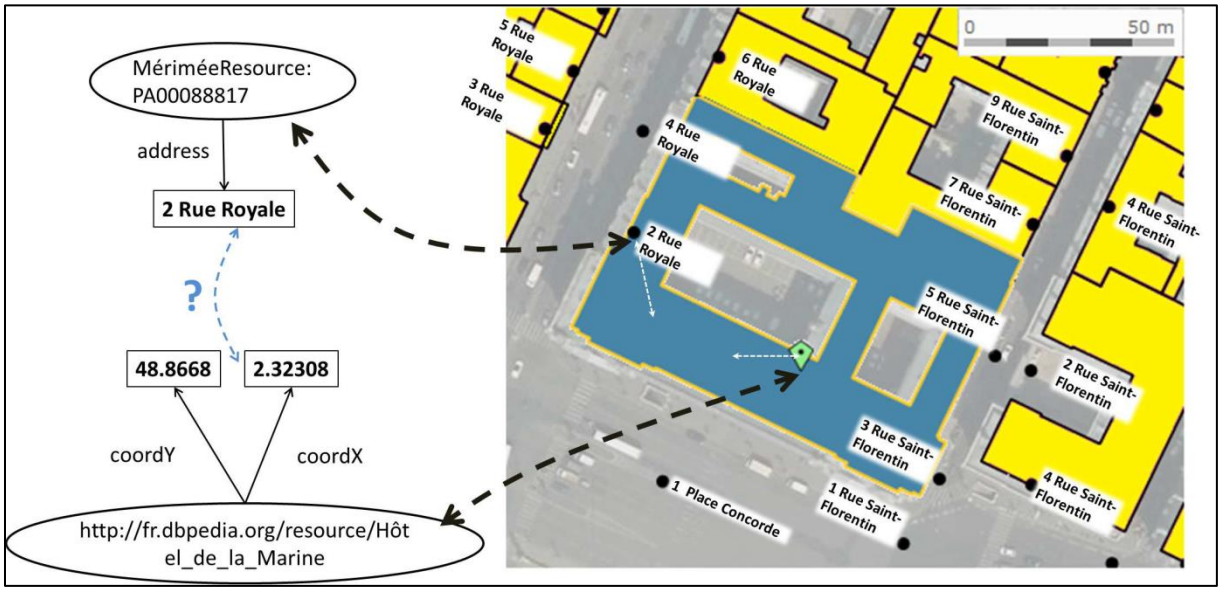


Figure 4.17 Exemple de l'utilisation de données topographique de référence comme support pour l'interconnexion de ressources géoréférencées dans le Web

Cette approche part du constat que les données géographiques de référence fournissent une description des géométries beaucoup plus détaillée que celle fournie par les références spatiales des ressources dans le Web comme on peut le constater avec l'exemple du bâtiment abritant l'hôtel de la marine dans la Figure 4.17.

Ancrer les ressources aux données géographiques de référence revient à réaliser une phase d'interconnexion qui utilise la comparaison des géométries comme critère principal de mise en correspondance. L'approche classique naïve, qui lie chaque ressource d'un jeu de données source à la ressource la plus proche spatialement dans le jeu de données cible, peut être appliquée dans ce

cas. Les outils d'interconnexion classiques peuvent donc être employés dans cette phase. Toutefois, une hypothèse principale conditionne cette phase d'ancrage : les ressources à interconnecter (que nous appelons ici ressources thématiques[117]) ne doivent être ancrées qu'à des données topographiques de référence qui sont sémantiquement cohérentes avec leur type. Par exemple s'il s'agit de ressources thématiques qui décrivent des musées, il est sémantiquement cohérent de les ancrer dans des données géographiques de référence qui représentent des bâtiments. Ou s'il s'agit de ressources thématiques décrivant des lacs, il est sémantiquement cohérent de les ancrer à des données géographiques de référence qui représentent des surfaces d'eau. Une deuxième hypothèse pour le bon déroulement de l'approche intervient quand il s'agit d'ancrer des données géoréférencées par des références spatiales indirectes (ex. adresse, toponyme, etc.) : Il faut disposer de données de référence qui permettent de géocoder ces références spatiales indirectes.

Dériver les liens d'interconnexion entre deux sources de données thématiques ancrées à un jeu de données géographiques de référence revient à croiser les deux ensembles de liens d'ancrage calculés pour les deux sources thématiques. Ceci s'apparente au fait de calculer des liens d'interconnexion par transitivité, et peut être facilement réalisé grâce à des requêtes Sparql. Lorsque deux ressources thématiques de 2 jeux de données différents sont liées à une même ressource topographique de référence, alors on crée un lien de correspondance entre ces deux ressources. La raison pour laquelle des liens de cardinalité n:m peuvent être créés dans ce cas est que plusieurs ressources d'une même source de données thématiques peuvent être ancrées au même objet topographique de référence. Cette phase de dérivation ouvre la porte à d'autres critères de création de liens. On peut éventuellement rajouter une vérification d'autres propriétés, telles que le nom, avant de décider si un lien doit être dérivé entre deux ressources thématiques ancrées à une même ressource géographique de référence.

Nous proposons dans la suite une mise en œuvre de cette approche appliquée au cas des monuments historiques de Paris présenté dans 4.1.4. En effet, nous avons constaté que certaines ressources de type monument dans une source de données peuvent être détaillées en plusieurs ressources de type monuments dans l'autre source. Nous démontrons avec cet exemple comment on peut détecter des liens d'interconnexion de cardinalité n:m d'une manière beaucoup plus précise qu'avec l'approche classique.

4.2.2 Mise en œuvre et évaluation de l'approche pour l'interconnexion des monuments parisiens

Nous avons repris l'exemple des deux jeux de données de monuments parisiens DBpedia et Mérimée présenté dans 4.1.4. Comme jeux de données topographiques de référence, nous avons utilisé deux composantes du référentiel à grande échelle (RGE) de l'IGN : les adresses de la BD ADRESSE® (présentée dans l'exemple de la section 3.2.1) et les bâtiments de la BD PARECELLAIRE® (présentée dans l'exemple de mise en œuvre de la section 3.2.2). Nous avons utilisé Silk comme un moyen d'interconnexion pour l'ancrage des monuments aux données de référence.

[117] Par opposition aux données topographiques de référence.

Description des interconnexions réalisées

La Figure 4.18 résume les différentes étapes de l'approche. Comme présenté dans 4.1.4, toutes les données sont transformées dans le modèle RDF grâce à la plateforme Datalift.

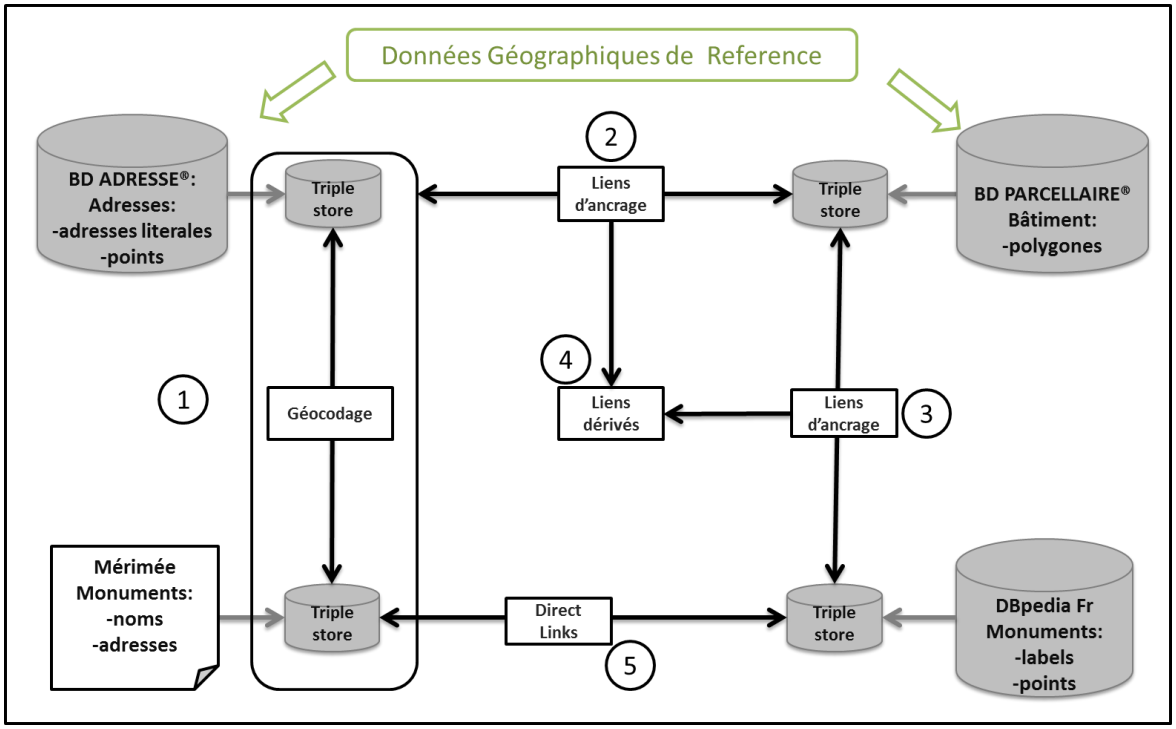


Figure 4.18 Etapes effectuées dans le cadre de l'approche d'interconnexion par ancrage-dérivation basée sur les données topographiques de référence.

Les ressources décrivant les monuments Mérimée ont été géocodage par interconnexion (Figure 4.18 (1)) comme nous l'avons expliqué dans 4.1.4. Les deux jeux de données de référence sont cohérents géométriquement. Les monuments Mérimée sont ancrés aux bâtiments de la BD PARCELLAIRE® en effectuant une étape d'interconnexion en utilisant les géométries récupérées de la BD ADRESSE® (Figure 4.18 (2)). Cette interconnexion lie chaque point adresse au bâtiment le plus proche en utilisant seulement des comparaisons géométriques avec un seuil de distance de 40m. Une interconnexion semblable est appliquée dans un second temps (Figure 4.18 (3)) pour ancrer les ressources DBpedia décrivant des monuments aux ressources qui décrivent les bâtiments de la BD PARCELLAIRE®. Dans les deux cas, la comparaison géométrique est effectuée en utilisant la mesure que nous avons rajoutée à Silk pour calculer la distance entre un point et n'importe quelle géométrie (*cf.* Figure 4.8). La dérivation des liens (Figure 4.18 (4)) est effectuée ensuite en créant des liens entre les ressources de type monument ancrées à la même ressource de type bâtiment.

Résultats et discussion

Afin de démontrer l'apport de cette approche, les résultats d'interconnexion résultant de la dérivation ont été rajoutés aux résultats obtenus par l'approche adaptative (*cf.* section 4.1). La somme des résultats des deux approches est comparée aux résultats d'une interconnexion directe des données. Cette interconnexion directe (Figure 4.18 (5)) ressemble à celle effectuée avec la méthode classique (tests comparatifs de la section 4.1.4). La différence majeure dans ce cas réside dans la recherche de liens de cardinalités n:m au lieu de 1:1. Pour réaliser cette interconnexion, nous

avons choisi un seuil de distance de 40m qui correspond au résultat de référence obtenu dans le cas de l'approche classique présentée dans les tests comparatifs de la section 4.1.4. Dans ce cas, il n'est pas nécessaire de fusionner l'ensemble des liens résultant pour une cardinalité 1:1 avec l'ensemble des liens obtenus pour une cordialité n:m, puisque le premier ensemble est forcément inclus dans le deuxième.

L'ensemble des liens obtenus par l'approche d'ancrage-dérivation est fusionné avec l'ensemble des liens obtenus par l'approche adaptative (en utilisant à la base de règles BR_2 dans les tests comparatifs de la section 4.1.4). Le Tableau 4.4 et la Figure 4.19 présentent les résultats obtenus par nos approches est les compare à ceux obtenus avec l'approche classique qui cherche des cardinalités n:m.

approches	Précision	Rappel	F-mesure
Approche classique n:m (seuil= 40m)	39,52%	58,88%	47,29%
Approche adaptative+ Approche d'ancrage-dérivation	**63,03%**	**63,60%**	**63,31%**

Tableau 4.4 Évaluation des résultats d'interconnexion globaux obtenus par nos approches par rapport aux résultats obtenus par l'approche classique.

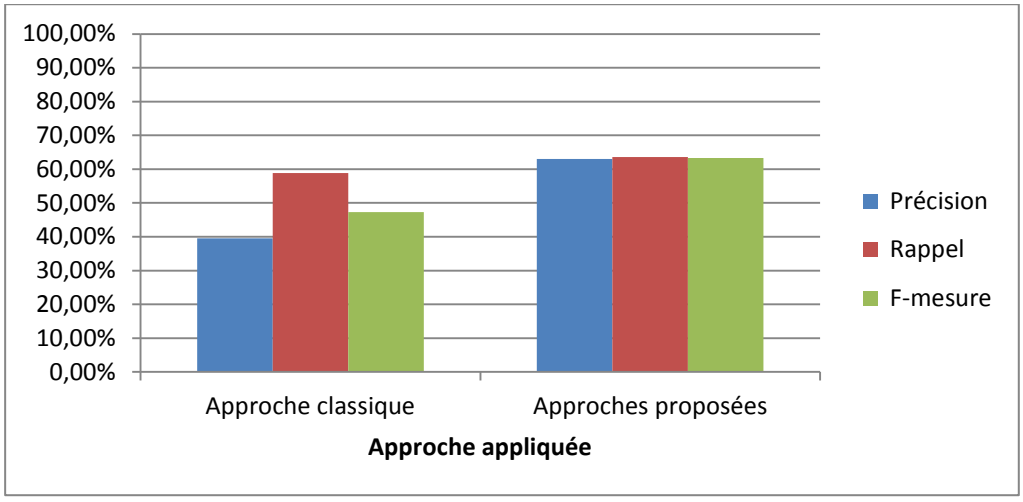


Figure 4.19 Résultats d'interconnexion globaux obtenus par nos approches par rapport aux résultats obtenus par l'approche classique

Les résultats obtenus par nos approches sont nettement meilleurs que ceux obtenu par l'approche classique. La recherche de lien de cardinalités n:m par l'approche classique permet d'augmenter le nombre de liens découverts sans pour autant améliorer le rappel. Ceci signifie que tous les nouveaux liens découverts ici sont faux par rapport à ceux découverts par l'approche classique qui cherche des liens 1:1. Ceci engendre une perte considérable en précision des résultats.

L'utilisation des données de référence comme support pour l'interconnexion montre principalement deux comportements positifs : elle permet d'augmenter le rappel en découvrant de nouveaux liens de cardinalité n:m (*ex.* Figure 4.20) ou des liens entre des ressources dont l'écart entre les géométries est très important (*ex.* Figure 4.21), tout en réduisant la perte en précision en évitant d'interconnecter des monuments qui sont spatialement proches, mais qui sont distincts (*ex.* Figure 4.22).

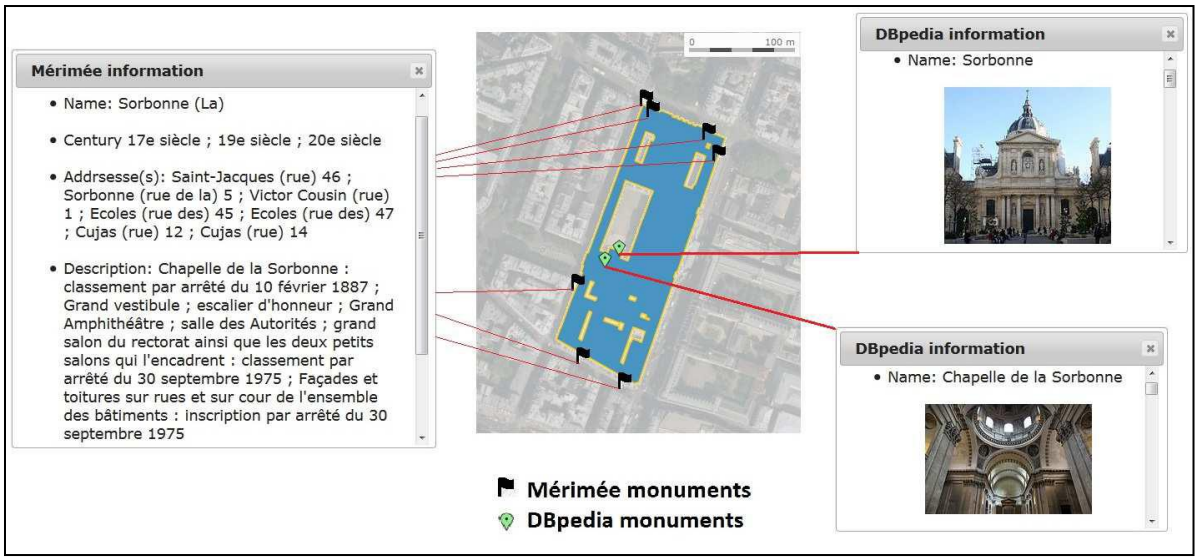


Figure 4.20 Exemple de liens de cardinalité n:m découvert par notre approche d'ancrage-dérivation.

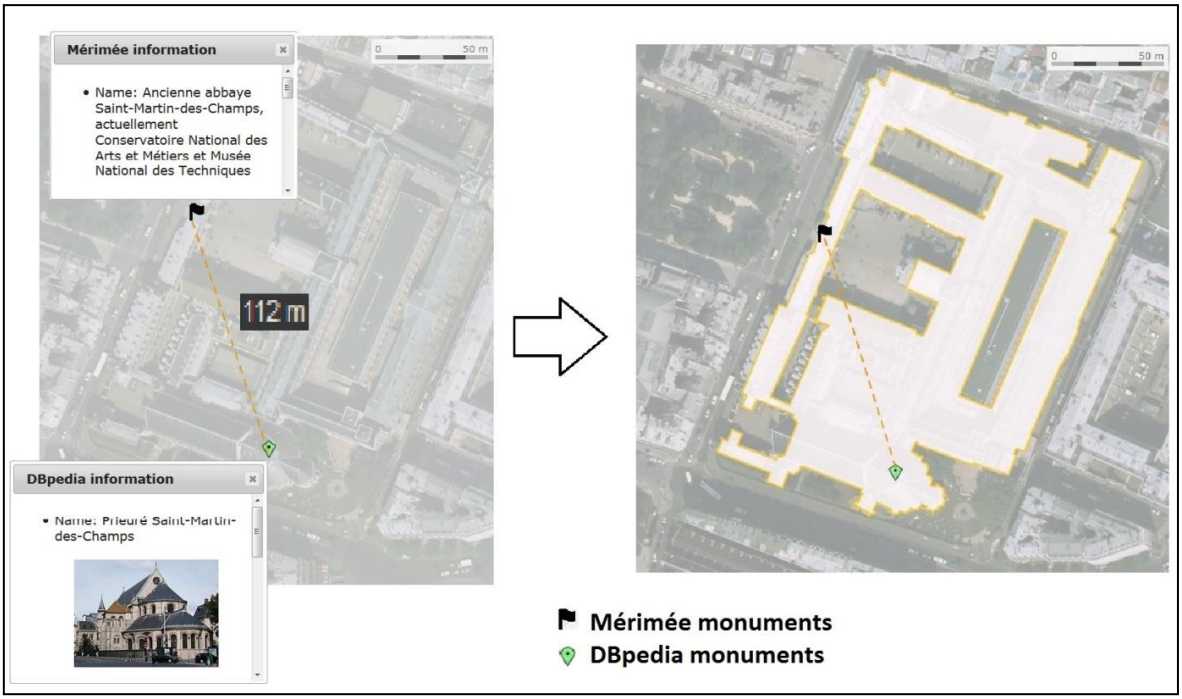


Figure 4.21 Exemple de nouveau lien découvert entre deux ressources très distantes spatialement grâce à notre approche d'ancrage-dérivation.

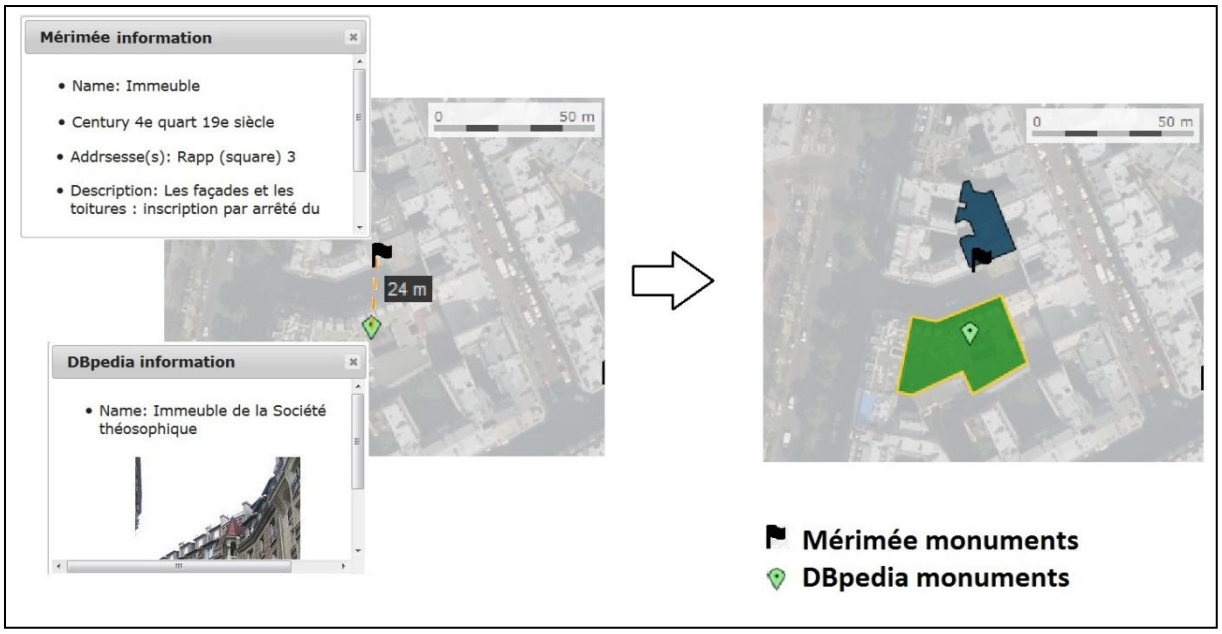


Figure 4.22 Exemple de faux lien découvert par l'approche classique, mais évité par notre approche d'ancrage-dérivation. Ceci est dû au fait d'ancrer des monuments proches mais distincts à des bâtiments différents.

Le gain en rappel et la réduction de perte en précision apportés par cette approche par rapport à l'approche classique sont obtenus grâce à certaines propriétés des données topographiques de référence : les géométries bien détaillées, la cohérence et l'homogénéité des données permettent de réduire l'effet des hétérogénéités des références spatiales des données à interconnecter.

Bien que les caractéristiques des géométries des ressources à interconnecter ne soient pas utilisées d'une manière explicite dans l'interconnexion comme c'est le cas avec l'approche adaptative, elles restent toutefois importantes pour la mise en œuvre de cette approche. En effet, le choix des données topographique de référence dépend de la modalisation géométrique des références spatiales des données à interconnecter. Nous avons choisi ici les géométries des bâtiments comme support pour l'interconnexion, car les métadonnées de modélisation géométrique des références spatiales indiquent principalement qu'elles sont saisies par rapport aux bâtiments ou aux voies routières qui donnent sur ces bâtiments.

4.3 Applications de Visualisation cartographique de données thématiques géoréférencées

Dans cette partie nous nous intéressons à l'une des applications les plus directement liées aux données géoréférencées et leur interconnexion, à savoir la visualisation cartographique. Plusieurs applications de visualisation cartographique de données géoréférencées ont été proposées dans le cadre du Web de données. Nous avons présenté dans la section 1.5 quelques exemples d'applications existantes qui utilisent les références spatiales comme un moyen de visualisation. Nous proposons ici deux applications qui s'appuient sur les références spatiales ainsi que les liens d'interconnexion entre données thématiques et données géographiques de référence pour la visualisation d'informations thématiques. Pour cela, nous nous appuyons sur les du domaine de la cartographie Web. Nous proposons dans un premier temps une application cartographique liée à un jeu de données spécifique en reprenant l'exemple des monuments historiques traité dans la section

4.2.2. Nous proposons ensuite une application plus générique d'exploration de données conçue pour plus de flexibilité dans la visualisation d'informations thématiques.

4.3.1 Exploitation des liens d'ancrage entre données thématiques et données topographiques de support pour la visualisation multi-échelle

Nous avons vu que l'approche d'interconnexion proposée dans la section 4.2.1 s'appuie sur une première étape d'ancrage. Cette étape permet de créer des liens explicites entre des ressources thématiques géoréférencées et des ressources topographiques issues d'un jeu de données de référence. En plus de constituer une source de connaissances de support pour l'interconnexion, nous considérons que ces données topographiques de référence peuvent fournir un support intéressant pour la visualisation. Au lieu de projeter les points qui géoréférencent les ressources thématiques du Web, nous proposons d'utiliser les géométries des données topographiques auxquelles elles sont liées comme moyen de visualisation. Ceci permet de disposer de géométries dotées d'un niveau de détail plus adapté à une visualisation à grande échelle.

Nous proposons une solution d'application Web légère qui s'appuie sur les liens d'interconnexion afin de fournir une visualisation cartographique multi-échelle d'informations thématiques. La Figure 4.23 résume l'architecture de cette application. Il s'agit d'une interface Web interactive, implémentée en utilisant les bibliothèques de cartographie Web Openlayers2[118] et l'API Géoportail[119]. Ces deux bibliothèques permettent de créer des couches de données vecteur à partir de données distantes. Nous avons rajouté une méthode qui permet d'interroger et récupérer des données à partir d'un point d'accès Sparql par des requêtes HTTP. Ainsi, l'application récupère les géométries des ressources topographiques de référence stockées dans un triplestore séparé des données thématiques. Grâce aux liens d'ancrage créés auparavant entre données thématiques et données topographiques, nous pouvons également récupérer l'information thématique et l'associer aux géométries des ressources topographiques correspondantes. La bibliothèque de l'API Géoportail permet également de récupérer et visualiser des fonds cartographiques et orthophotographiques en flux à partir du serveur Géoportail[120].

Nous avons utilisé cette application de visualisation sur le cas des monuments historiques de Paris ancrés aux bâtiments de la BD PARECELLAIRE®, présenté dans la section 4.2.2. Comme le montre la Figure 4.24, nous arrivons grâce aux liens d'ancrage à co-visualiser les informations sur les monuments historiques issus des deux sources DBpedia et Mérimée. Dans cette figure, les polygones ayant un contour orangé représentent des bâtiments liés à des monuments DBpedia. Les polygones teintés en bleu représentent des bâtiments liés à des monuments Mérimée. La graduation du bleu représente ici le siècle de construction du monument historique (entre le 10ème et le 20ème siècle) renseigné par l'attribut « siècle » dans la description des monuments Mérimée. La couleur noire signifie que le siècle de construction du monument n'a pas été renseigné. L'un des avantages de cette architecture est de pouvoir récupérer à la volée par des requêtes HTTP, grâce aux liens d'ancrage, des informations thématiques issues des descriptions des ressources DBpedia et Mérimée en cliquant sur la géométrie d'un bâtiment.

[118] http://openlayers.org/
[119] http://api.ign.fr/
[120] https://www.geoportail.gouv.fr/

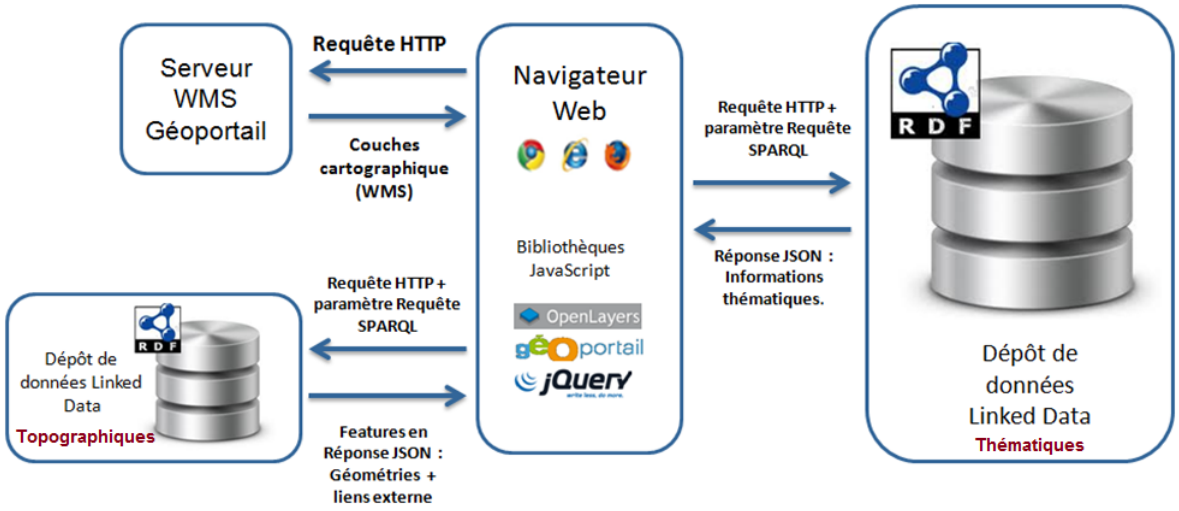


Figure 4.23 Architecture de l'application de visualisation cartographique des données thématique grâce à leur interconnexion avec des données topographiques de référence.

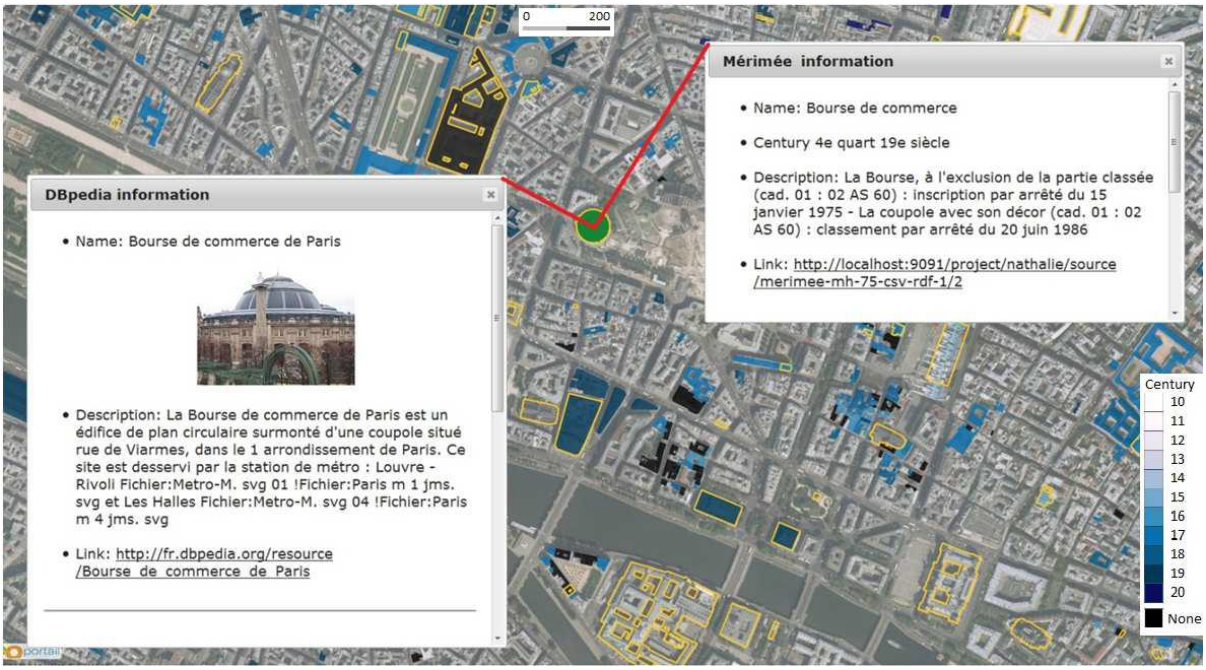


Figure 4.24 Co-visualisation des données thématiques des deux sources Mérimée et DBpedia à grande échelle grâce aux liens d'ancrage avec les bâtiments des données topographiques de support BD PARCELLAIRE®

La visualisation que nous proposons dans la Figure 4.24 est mieux adaptée à grande échelle, car elle affiche les bâtiments à un niveau de détail géométrique élevé. À petite échelle les bâtiments deviennent difficiles à discerner individuellement, ce qui nécessite une adaptation du contenu affiché. Des opérations de généralisation sont dans ce cas nécessaires pour générer une cartographie claire et lisible. En effet, le but de la généralisation est de simplifier le contenu affiché quand on change l'échelle tout en gardant le sens de la carte (Ruas, 2002).

Nous nous appuyons donc sur les approches existantes dans le domaine de la généralisation cartographique afin de proposer une solution de visualisation multi-échelle de l'information thématique. Nous utilisons les polygones qui décrivent les bâtiments auxquels nous avons ancré les

monuments Mérimée comme moyen d'agrégation. En effet, les bâtiments sont regroupés et amalgamés selon leur proximité géométrique et leur information thématique associée. Pour cela, nous avons adapté l'algorithme d'amalgamation de blocs (clusters) de polygones proposé par (Regnauld, 2003). Nous proposons un algorithme basé sur trois étapes principales :

- **Le regroupement** : Cette étape consiste à créer des groupes de géométries qui sont spatialement proches et qui partagent une même information thématique. La proximité spatiale est vérifiée en créant des zones tampons d'un rayon r autour des polygones et en cherchant toutes les intersections possibles. Vérifier le partage de la même information thématique revient ici à vérifier si les bâtiments sont liés à des ressources ayant dans leur description une même valeur de propriété, ici le siècle de construction.
- **L'amalgamation** : Dans cette étape, les groupes de géométries créés à l'étape précédente sont traités séparément. Les géométries de chaque groupe sont amalgamées en une seule géométrie dont la visualisation a un effet visuel proche de celui de la visualisation indépendante des bâtiments. Pour cela, nous avons utilisé un algorithme proposé par (Duckham et al., 2008) basé sur la triangulation de Delaunay pour la création d'une enveloppe concave à partir d'un ensemble de points. Nous avons réutilisé une implémentation[121] Java existante de cet algorithme. Cet algorithme étant conçu principalement pour fonctionner avec un ensemble des points, son comportement avec des polygones n'est pas toujours optimal. Pour pallier ce problème, nous commençons par une densification des segments des polygones avant de procéder à la création de l'enveloppe concave. L'amalgamation d'un groupe de géométries résulte en une seule géométrie qui porte l'information thématique partagée par les monuments du groupe ainsi que des liens vers les ressources qui décrivent ces monuments.
- **Le filtrage :** Cette étape consiste à enlever de l'affichage toute géométrie non pertinente par rapport à l'échelle de visualisation. Le but de l'amalgamation étant de créer des objets suffisamment grands pour être lisible sur la carte, nous écartons tout objet ayant une superficie inférieure à un seuil a. Cette étape sert aussi à la gestion des superpositions des géométries amalgamées. En effet, les géométries les plus grandes sont prioritairement mises en avant, et si une géométrie est complètement recouverte par une autre ou que sa partie apparente à une superficie inférieure au seuil a, elle est supprimée de la visualisation.

Le résultat de cet algorithme est présenté dans la Figure 4.25. Dans cet exemple, nous avons choisi des valeurs de r et a adaptées à une échelle d'affichage au niveau du « quartier » (entre le niveau « rue » et le niveau « ville » du Géoportail).

Pour une échelle encore plus petite (niveau « ville » du Géoportail ou plus petite) nous proposons d'augmenter les valeurs de r et a , afin d'amplifier les géométries résultantes pour avoir une visualisation lisible (voir l'exemple de la Figure 4.26.)

[121] www.rotefabrik.free.fr/concave_hull/

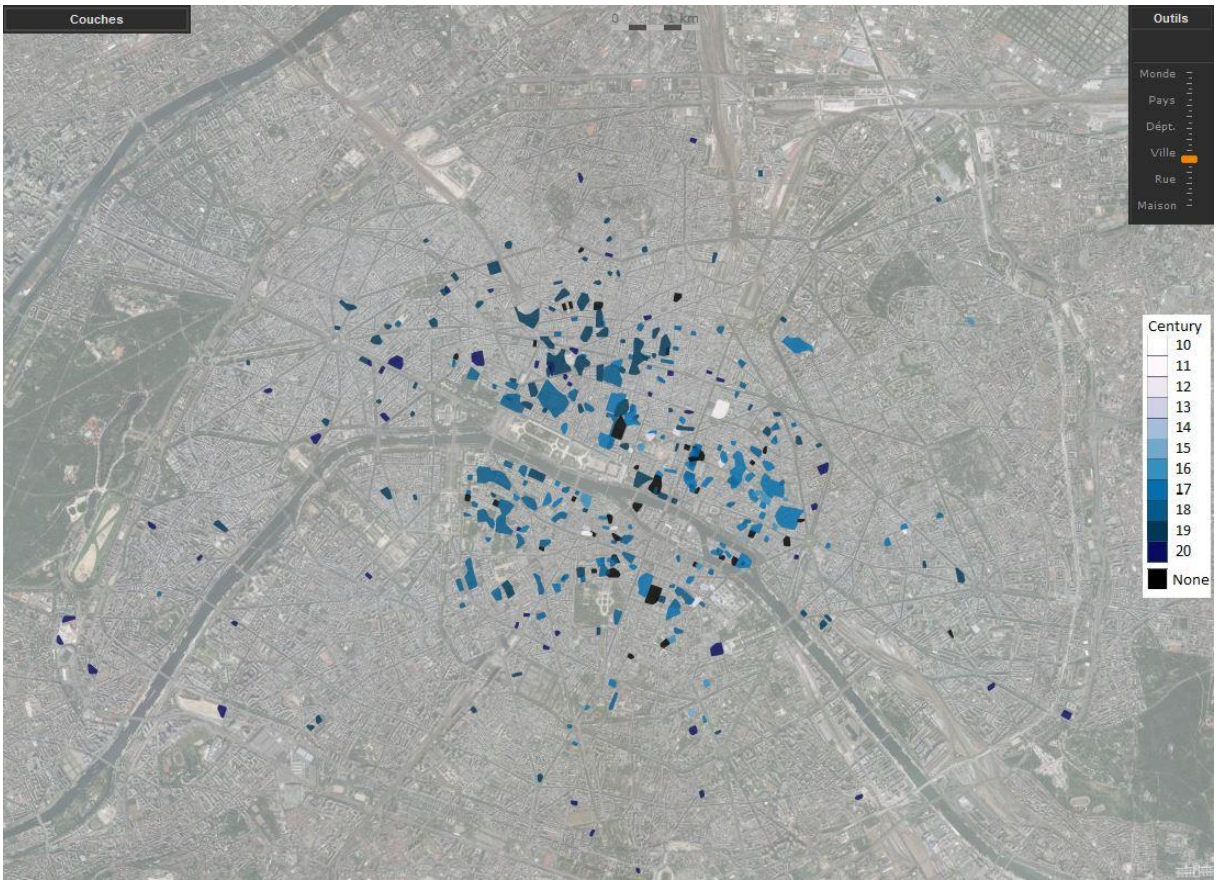


Figure 4.25 Résultats de l'application de l'algorithme de regroupement-amalgamation des géométries des bâtiments pour la visualisation des données thématiques Mérimée (siècles de construction des monuments) à l'échelle d'affichage des « quartiers » (entre les niveaux « rue » et « ville »).

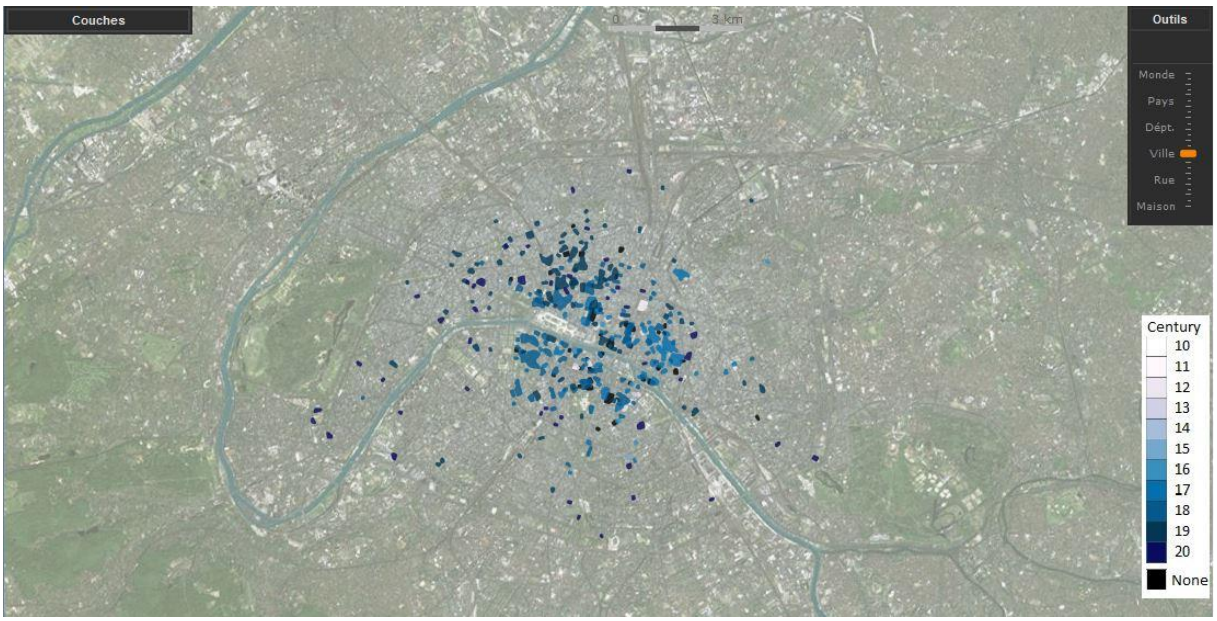


Figure 4.26 Résultats de l'application de l'algorithme de regroupement-amalgamation des géométries des bâtiments pour la visualisation des données thématiques Mérimée (siècles de construction des monuments) à l'échelle d'affichage de la « ville » ou plus petite.

L'application de visualisation multi-échelle que nous avons présentée ici se distingue clairement des solutions de visualisation de données liées existantes car elle permet de fournir une visualisation à des échelles différentes en gardant un contenu à la fois cohérent et lisible entre ces échelles : elle permet dans, cet exemple, d'avoir une idée sur la distribution des âges des monuments dans la ville. Cette application est un prototype qui vise à démontrer le potentiel de l'utilisation des liens d'interconnexion entre données thématiques et données topographiques de référence dans la création d'applications de visualisation cartographique plus conviviales, plus interactives et plus lisibles. Elle reste cependant limitée car elle n'est implémentée que pour le cas des monuments parisiens. En outre, l'algorithme d'amalgamation que nous proposons dans ce prototype n'est pas appliqué à la volée sur les données : les géométries, telles qu'elles sont représentées pour les petites échelles (Figure 4.25 et Figure 4.26) sont pré-calculées et récupérées lorsqu'un changement de l'échelle de visualisation survient. Finalement, l'architecture à base de client léger, où l'interrogation des points d'accès Sparql et les traitements de création de couches vectorielles à afficher s'effectuent du côté de l'application Web cliente, n'est pas adaptée à des cas de visualisation de données volumineuses. Ces problématiques nous ont poussé à envisager une application plus générique d'exploration de sources de données liées géoréférencées.

4.3.2 Vers une application d'exploration cartographique multi-échelle générique des sources de données géoréférencées

Afin de proposer une application d'exploration et de visualisation cartographique multi-échelle de jeux de données liées géoréférencées, nous avons analysé les différentes applications de visualisation cartographique de données sur le Web, en particulier celles conçues pour des données liées. Ce qui ressort de cette analyse sur des applications de visualisation cartographique des données liées, telles que ceux présentés dans la partie 1.5, est l'absence de gestion de la représentation multi-échelle. Cependant, d'une manière plus générale, les solutions de visualisation cartographique des données sur le Web utilisent parfois des techniques pour fournir des représentations multi-échelles lisibles. Nous avons identifié principalement deux techniques utilisées pour réduire la quantité d'entités géographiques à afficher à petite échelle : la sélection et l'agrégation.

Techniques existantes pour la visualisation multi-échelle des données sur le Web

La sélection consiste à n'afficher à l'écran qu'un nombre fini de valeurs, souvent les plus importantes. L'affichage des valeurs les moins significatifs se fait lorsque l'échelle d'affichage augmente. Ce type de méthode s'applique pour des données du type « quantitative absolue » ou « qualitative ordinale ». L'avantage principal de cette technique réside dans sa simplicité de mise en œuvre quand on change l'échelle de visualisation. Le principal inconvénient réside dans la non-présence de toutes les données sur la carte ce qui peut induire en erreur dans certains cas (l'absence de représentation n'est pas synonyme d'absence d'information). La Figure 4.27 montre deux exemples de solutions de visualisation de données qui utilisent la technique de sélection pour l'affichage multi-échelle. La partie gauche de cette figure correspond au site Géoclip[122] qui utilise une technique de sélection avec un seuil de coupe de représentation des données qui permet de voir un contexte général et des

[122] http://franceo3.geoclip.fr/

contextes locaux dans la répartition de la population en France. La partie droite de la figure correspond à la visualisation cartographique des différents hôtels, locations de vacances, restaurants et activités de la ville de Bordeaux (données qualitatives ordonnées suivant la moyenne des notes attribuées) sur le site Tripadvisor[123]. Ce site utilise une technique de sélection favorisant les valeurs dominantes, qui sont représentées avec une taille plus forte, ce qui permet de faire ressortir parmi l'ensemble l'importance de ces valeurs.

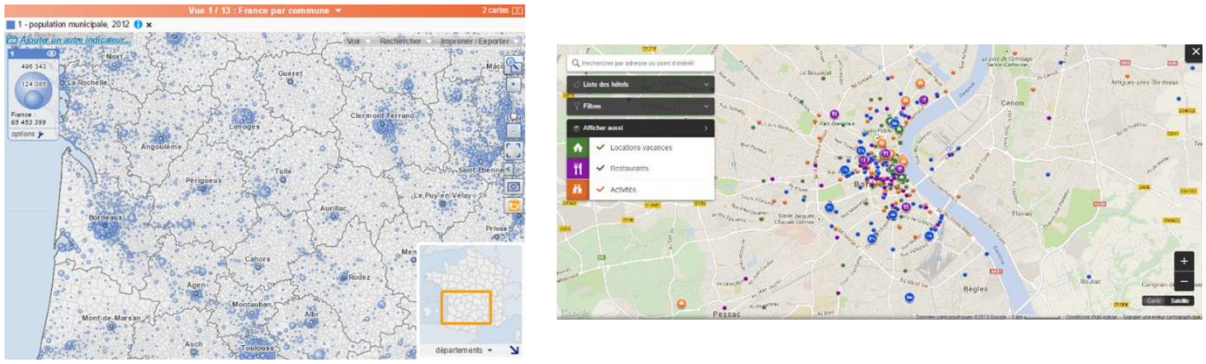


Figure 4.27 Solution de visualisation utilisant des techniques de sélection. À gauche une carte représentant la population par communes sur le site Géoclip. À droite une carte de localisation des sites touristique sur le site Tripadvisor.

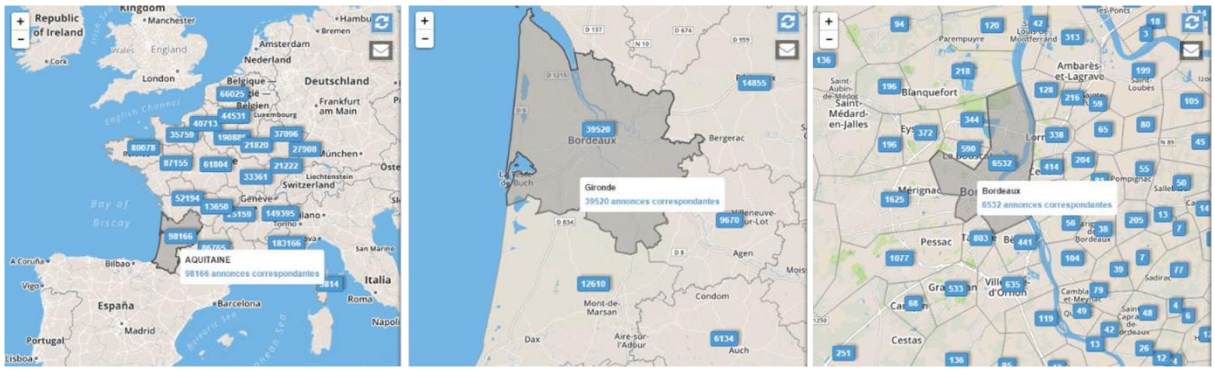


Figure 4.28 Cartes répertoriant les offres immobilières en France sur le site Homengo : exemple d'agrégation par comptage.

L'agrégation est une technique qui consiste à fusionner, combiner, additionner, moyenner,etc. des données, lorsque les niveaux de zoom deviennent trop petits ou les informations trop denses. Parmi la panoplie des méthodes d'agrégation possibles, le compteur (ou *buffer*) semble être la plus répandue. L'avantage principal de cette technique réside dans la simplification de l'information représentée dès que l'échelle d'affichage le permet. Le principal inconvénient est que la lecture de la carte n'est pas toujours simple, car elle dépend du type d'agrégation effectuée. Pour que la carte puisse être comprise, la méthode d'agrégation doit être connue. La plupart du temps cette méthode nécessite d'utiliser des zonages pour les différents niveaux d'échelle. Le choix du zonage a une forte importance sur le résultat d'agrégation. La Figure 4.28 montre un exemple de visualisation cartographique des données du site Homengo[124]. En utilisant une technique d'agrégation par comptage, les cartes présentées ici répertorient le nombre d'offres immobilières en France.

[123] http://www.tripadvisor.fr/
[124] https://homengo.com/

L'utilisation d'un zonage différent et adapté pour chaque échelle d'affichage permet de se rendre compte de l'état de la situation immobilière en France depuis différents points de vue.

Proposition d'une application d'exploration cartographique multi-échelle de données thématiques

Nous traitons ici le cas de données thématiques liées qui sont géoréférencées par des géométries ponctuelles. Nous suggérons d'utiliser des données de zonage comme moyen pour agréger l'information thématique portées par ces données. L'application que nous proposons doit permettre l'exploration d'une source de données thématiques géoréférencées tout en laissant, d'une façon interactive, le choix à l'utilisateur de sélectionner l'information thématique à visualiser et sa nature, ainsi que la technique convenable (sélection ou agrégation) à utiliser pour la visualisation multi-échelle. Nous proposons de nous appuyer sur le vocabulaire utilisé pour structurer les données thématiques pour développer une approche d'agrégation générique de données. L'utilisateur doit répondre aux questions suivantes pour définir le rendu cartographique souhaité :

- Sur la nature des données :
- Quelle est l'information thématique à visualiser ?
- S'agit-il de données quantitatives ou qualitatives ?
- S'agit-il de données relatives ou absolues ? Ou nominales ou ordinales ?
- Si ce sont des données qualitatives nominales : sont-elles classées ou non classées ?
- Sur le choix de visualisation :
- Voulez-vous une visualisation statique ou dynamique de l'information ? (Dynamique dans le sens : Voulez-vous que le contenu de la carte soit adapté dynamiquement à l'échelle de visualisation utilisées?)
- Si les données sont qualitatives : Voulez-vous représenter un seul, plusieurs ou tous les éléments?

<table>
<tr><th colspan="3">Nature des données</th><th colspan="3">Données cartographiés</th></tr>
<tr><td rowspan="2">Quantitative</td><td colspan="2">Relative (Taux)</td><td colspan="2">Elément</td><td>Tous</td></tr>
<tr><td colspan="2">Absolue (Stock)</td><td colspan="2">Elément</td><td>Tous</td></tr>
<tr><td rowspan="6">Qualitative</td><td rowspan="4">Nominale</td><td rowspan="3">Classé</td><td rowspan="2">Statique</td><td rowspan="2">Nœud (feuille compris)</td><td>Tous</td></tr>
<tr><td>Plusieurs</td></tr>
<tr><td>Un seul</td></tr>
<tr><td>Dyn.</td><td>Feuille * -> Nœud</td><td>Tous</td></tr>
<tr><td rowspan="3">Non classé</td><td colspan="2" rowspan="3">Elément</td><td>Tous</td></tr>
<tr><td>Plusieurs</td></tr>
<tr><td>Un seul</td></tr>
<tr><td rowspan="3">Ordinale</td><td rowspan="2">Statique</td><td rowspan="2">Ordre</td><td>Tous</td></tr>
<tr><td>Plusieurs</td></tr>
<tr><td>Un seul</td></tr>
<tr><td>Dyn.</td><td>Ordre -> Ordre</td><td>Tous</td></tr>
</table>

Figure 4.29 Arbre de décision sur les possibilités de cartographie des données selon leur nature.

Cette série de questions permet de parcourir l'arbre de décision présenté dans la Figure 4.29, ce qui permet de proposer à l'utilisateur une cartographie des donnés adéquate, selon leur nature. Quand les données à cartographier se présentent sous la forme de valeurs qualitatives ordinales (hiérarchie directe) ou des valeurs qualitatives nominales classées (notamment dans des arborescences sémantiques de concepts), deux possibilités de visualisation s'offrent à l'utilisateur : dynamique ou statique. La visualisation statique consiste à choisir un seul niveau de valeurs à cartographier dans la hiérarchie ou dans la structure de classification. La visualisation dynamique consiste à changer dynamiquement, selon l'échelle d'affichage, les valeurs à cartographier en remplaçant les valeurs initiales par les valeurs plus générales correspondantes dans la hiérarchie des concepts proposée par le vocabulaire qui décrit les données. En outre, nous proposons de nous appuyer sur un maillage géographique pour agréger les informations thématique lorsque leur représentation individuelle n'est plus possible pour des raisons de lisibilité.

Cet arbre de décision permet à l'utilisateur de choisir des méthodes de visualisation multi-échelle telles que la sélection, le compteur, la moyenne, la densité, etc. Ceci permet donc de fournir une variété de visualisations possibles qui peuvent aller d'un simple affichage de points sur un fond cartographique jusqu'à la création de cartes statistiques.

Implémentation de l'approche

Nous proposons un prototype pour cette application de visualisation et d'exploration cartographique multi-échelle de données thématiques. Les données de maillage utilisées sont celles du découpage administratif (départements, communes…). Il s'agit de la base GEOFLA® de l'IGN déjà disponible en RDF[125]. Nous avons choisi un jeu de données thématique de test pour la mise au point de ce prototype. Il s'agit de la base du « Patrimoine Gourmand[126] » d'Ile de France disponible notamment au format Shapefile sous licence Ouverte. Cette base, composée de 3777 éléments géoréférencés par des coordonnées, est fournie avec un fichier de métadonnées qui explique son contenu. La Figure 4.30 montre un aperçu de ces données en les catégorisant selon leurs natures. Nous avons nettoyé et structuré cette base de données dans le modèle RDF. Les classes et propriétés utilisées sont définit dans une ontologie que nous avons définie en amont en analysant les métadonnées de cette base et les valeurs des différentes propriétés. Chaque site est décrit par des propriétés qui permettent de spécifier l'identifiant du site, sa dénomination, son siècle de construction, sa nature, le type de produit alimentaire qu'il fournit, etc.

[125] http://data.ign.fr/id/geofla/
[126] https://www.data.gouv.fr/fr/datasets/base-de-donnees-patrimoine-gourmand-d-ile-de-france-idf/

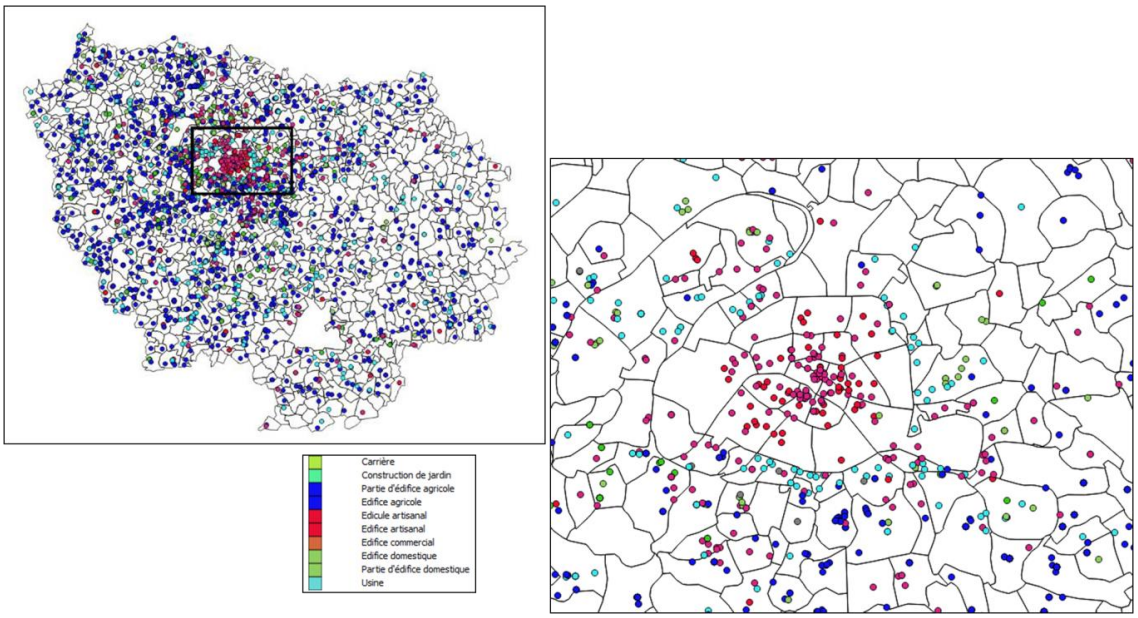


Figure 4.30 Aperçus et légende QGIS du jeu de données du « patrimoine gourmand » avec zoom sur Paris. Les couleurs représentent les différentes natures des sites représentés dans cette base.

Afin de dépasser les limites de l'architecture client-serveur de l'application de visualisation présentée dans la section 4.3.1, nous proposons ici une architecture 3-tiers (voir Figure 4.31).

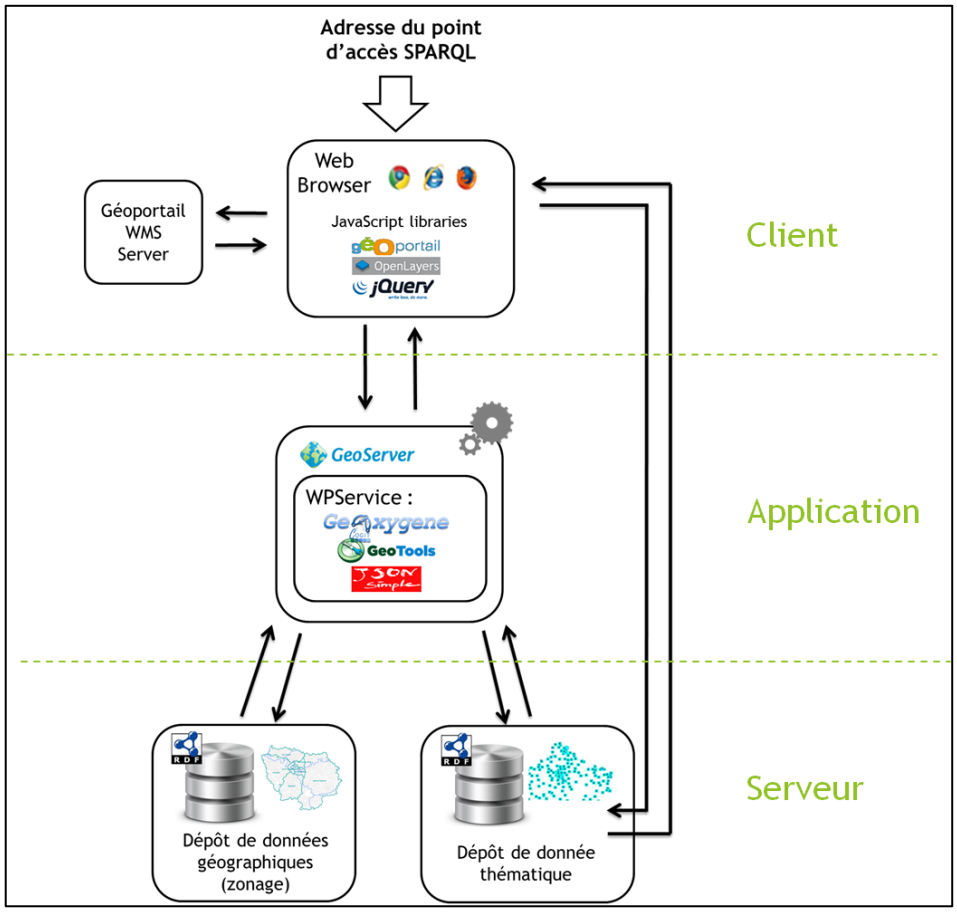


Figure 4.31 Architecture 3-tiers de l'application d'exploration cartographique multi-échelle .

Le niveau « serveur » comprend les *triplestores* dans lesquels les données thématiques et les données géographiques de maillage sont stockées. Le niveau « client » est constitué d'une interface Web fondée sur l'API Géoportail. Le niveau « application » permet d'alléger le traitement effectué par l'application cliente. Il se charge de récupérer les requêtes de la couche client et de les traduire sous forme de requêtes Sparql pour interroger les points d'accès de données. Il se charge ensuite de la création des couches à visualiser selon l'information thématiques sélectionnée par l'utilisateur ainsi que la méthode d'agrégation et le niveau de zoom choisis. Le niveau application est implémenté sous forme d'un service WPS (*Web Processing Service*) développé sous la plateforme GeOxygene[127] et installé dans un serveur cartographique Geoserver[128].

L'application gère la visualisation multi-échelle en associant un niveau du découpage administratif à chaque échelle de visualisation (en utilisant les niveaux de zoom définis dans l'API Géoportail). Ces associations sont présentées dans le Tableau 4.5.

Niveau de zoom	Maillage
4 et moins	Pays
5 et 7	Régions
8 et 9	Départements
10	Arrondissements
11	Cantons
12	Communes
13 et plus	Pas de zonage

Tableau 4.5 Association entre niveaux de zoom de visualisation et niveaux de données de zonage.

L'application présente un formulaire à l'utilisateur pour spécifier le point d'accès Sparql du jeu de données thématique à explorer et sélectionner l'information (la propriété) thématique à visualiser ainsi que sa nature et le choix de la technique de visualisation à utiliser (Figure 4.32). Dans ce formulaire on spécifie tout d'abord l'URI du point d'accès Sparql de la source de données à explorer. L'application récupère les URIs des différents graphes de données contenus dans cette source. En sélectionnant un graphe, l'application récupère à la volée ses différentes classes de ressources. En sélectionnant une classe spécifique, l'application récupère à la volée les différentes propriétés qui décrivent les instances de cette classe. En sélectionnant une propriété, l'application récupère des exemples de valeurs de cette propriété, qui permettent de mieux comprendre leur nature. L'utilisateur doit spécifier ensuite la nature de ces valeurs. L'application propose enfin les techniques de sélection ou d'agrégation adaptées à cette nature. Les informations sont envoyées de ce formulaire au niveau application qui crée des couches à visualiser à partir des données RDF fournies par le serveur et des informations transmises par le formulaire. Ces couches sont envoyées au niveau client pour les visualiser cartographiquement.

[127] http://ignf.github.io/geoxygene/index.html
[128] http://geoserver.org/

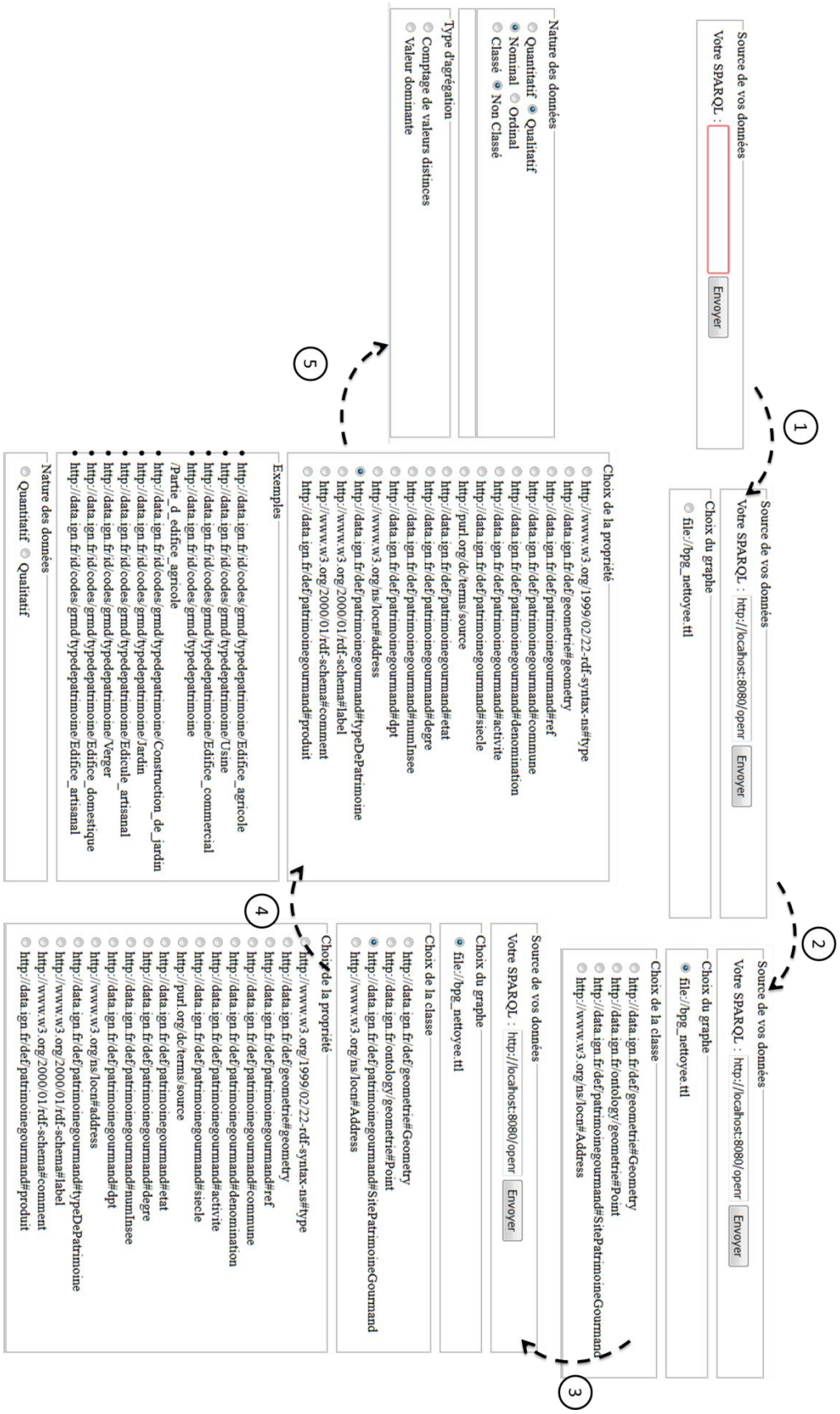


Figure 4.32 Formulaire d'interaction avec l'utilisateur pour le choix de l'information thématique, sa nature et la technique d'agrégation multi-échelle.

Nous avons implémenté quelques techniques d'agrégation pour tester le bon fonctionnement du prototype. La première technique consiste à agréger pour chaque maille de découpage administratif, les données en effectuant un comptage de valeurs distinctes de la propriété thématique à afficher. Ce comptage est représenté, au centre de chaque géométrie de zone, par un cercle de taille proportionnelle aux comptages minimum et maximum visible dans la fenêtre courante de visualisation. Chaque cercle proportionnel contient également le nombre de valeurs distinctes de la propriété visualisée comptabilisées dans la zone associée. La Figure 4.33 montre un exemple utilisant cette technique de comptage. Dans cet exemple, nous avons choisi l'identifiant des ressources (dont la valeur est unique pour chaque ressource) comme propriété thématique à visualiser. Le comptage permet donc ici d'afficher le nombre de ressources contenues dans chaque zone.

Une deuxième technique implémentée que l'utilisateur peut choisir pour la visualisation est l'agrégation des données en les représentant par la proportion de la valeur dominante d'une propréié données dans une zone par rapport au nombre total de ressources dans cette zone. Un aperçu de cette technique est présenté dans la Figure 4.34. Les diagrammes circulaires représentent la part de la valeur de propriété dominante dans à l'ensemble des valeurs du zonage. La taille des diagrammes est proportionnelle au nombre de ressources recensées dans le zonage. La propriété sélectionnée pour l'affichage dans ce cas est le « type » des ressources du patrimoine gourmand (édifice commercial, édifice artisanal, édifice agricole, etc.). En cliquant sur un diagramme, la valeur dominante de cette propriété dans le zonage correspondant est affichée. Ceci permet de déterminer si un type de site lié au patrimoine gourmand est plus courant que les autres dans une zone donnée, et le cas échéant identifier de quel type d'édifice il s'agit.

Les techniques proposées ici ont été appliquées d'une manière statique dans les deux exemples précédents, car il ne s'agit pas d'une propriété thématique classée ou ordonnée. Dans cet exemple de sites du patrimoine gourmand, nous avons proposé une taxonomie de concept SKOS des valeurs possibles de la propriété « type_produit » qui spécifie les types de produits fournis par chaque site (voir Figure 4.35), afin de tester la visualisation cartographique dynamique des valeurs qualitatives nominales classées. Nous avons appliqué la technique d'agrégation, présentée dans le paragraphe précédent, qui représente la proportion de la valeur dominante par rapport au nombre total de ressources dans une zone. Pour cette propriété, la technique est appliquée d'une manière dynamique. Pour permettre un changement dynamique du niveau hiérarchique des valeurs à afficher à partir de la taxonomie des types de produits, il est nécessaire d'associer automatiquement chaque niveau de maillage à un niveau dans la hiérarchie de la taxonomie des types de produits. L'idée est que le changement de visualisation d'un niveau de maillage plus détaillés au un niveau de maillage moins détaillé soit accompagner par un changement de la valeur de propriété visualisée d'un niveau moins général à niveau plus générale dans la taxonomie des valeurs de cette propriétés. L'association entre niveau de maillage et niveau dans la taxonomie doit donc prendre en compte le nombre total de niveaux de maillage N et le nombre total de niveaux dans la taxonomie des valeurs H (en excluant la racine si elle est unique). Pour chaque niveau de zonage d'ordre n , on associe le niveau hiérarchique dans une taxonomie h l'équation 4.6, tel que 6 est l'ordre du niveau de maillage « pays » et 1 est l'ordre du niveau de maillage « commune » et que 0 est l'ordre des concepts les moins générales (les feuilles) dans la taxonomie.

$$h = quotient\left(\frac{H \times (n-1)}{N}\right) \quad 4.6$$

Dans cet exemple la taxonomie est composée de niveau de concept. L'équation 4.6 a permis d'associer le niveau 0 de la taxonomie aux niveaux de maillege « commune », « canton » et « arrondissement », et le niveau 1 de la taxonomie aux niveaux de zonage « département », « région » et « pays » (voir Figure 4.35).

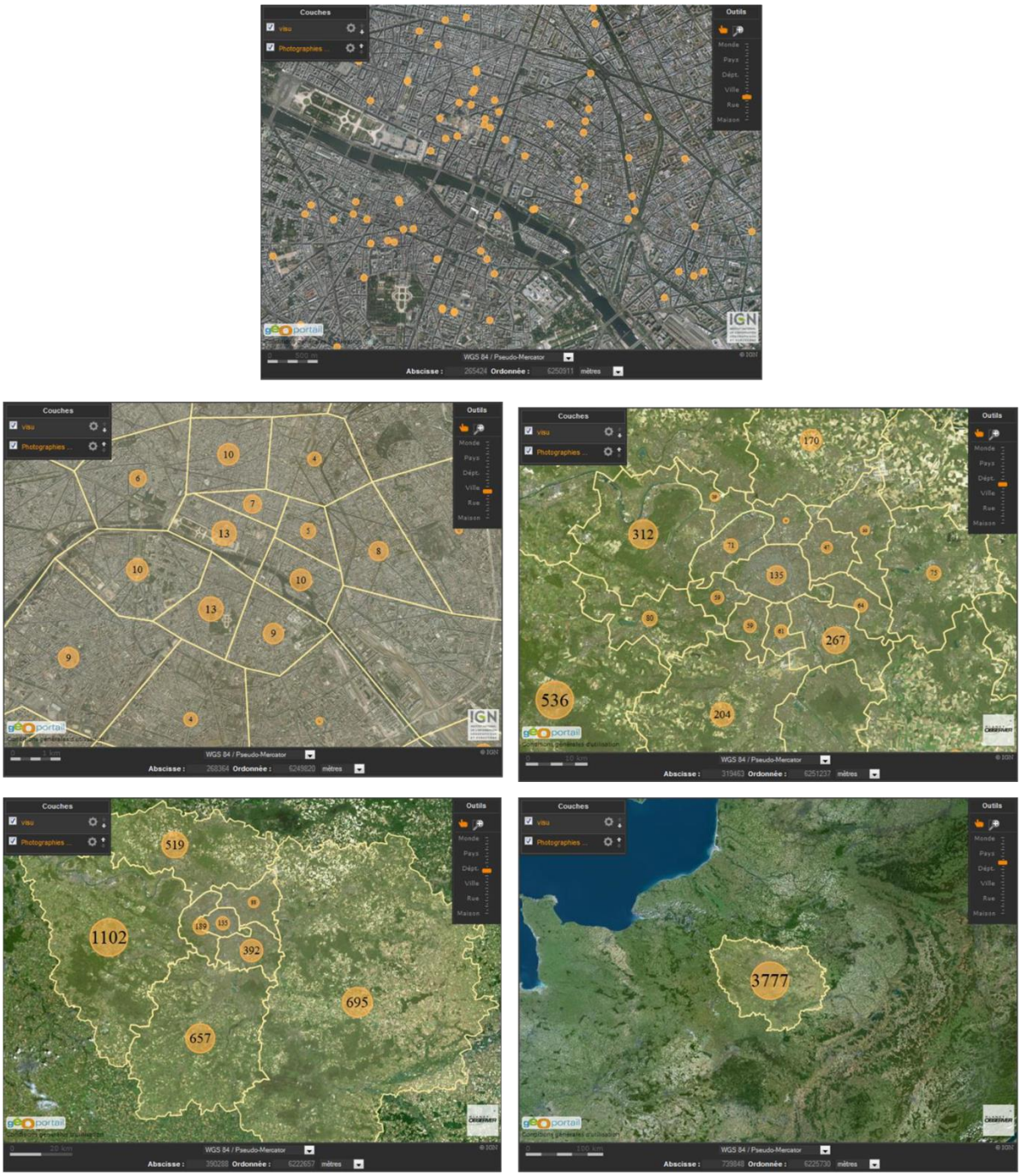


Figure 4.33 Utilisation de la technique d'agrégation par comptage de valeurs distinctes d'une propriété représentées dans des cercles proportionnels pour la visualisation multi-échelle de données thématiques. La propriété sélectionnée dans cet exemple est l'identifiant « ref » des ressources.

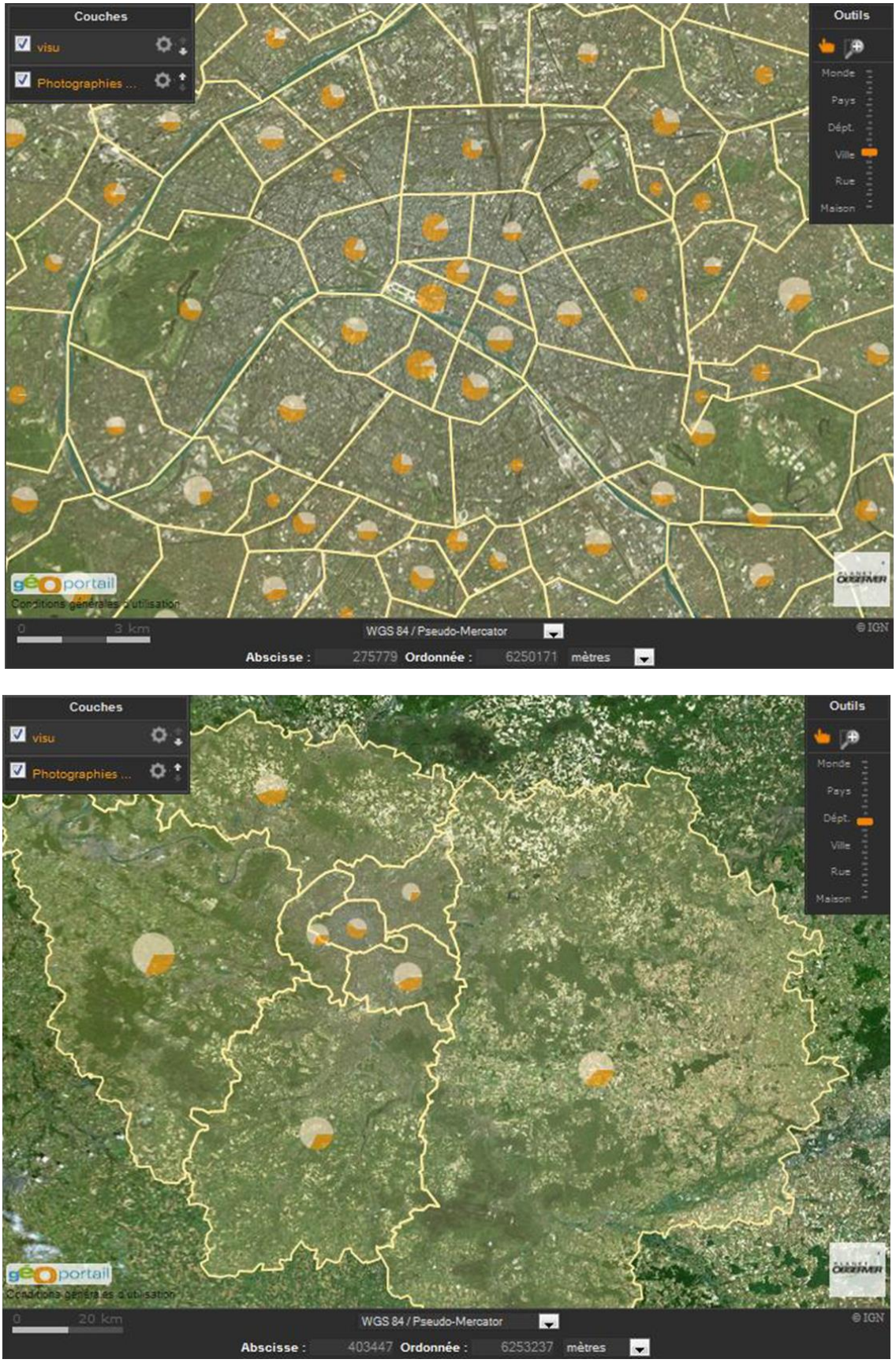


Figure 4.34 Utilisation de la technique d'agrégation par comptage des valeurs dominantes représentées par des diagrammes circulaires de tailles proportionnées pour la visualisation multi-échelle de données. La propriété sélectionnée dans ce cas est le type du site du patrimoine gourmand.

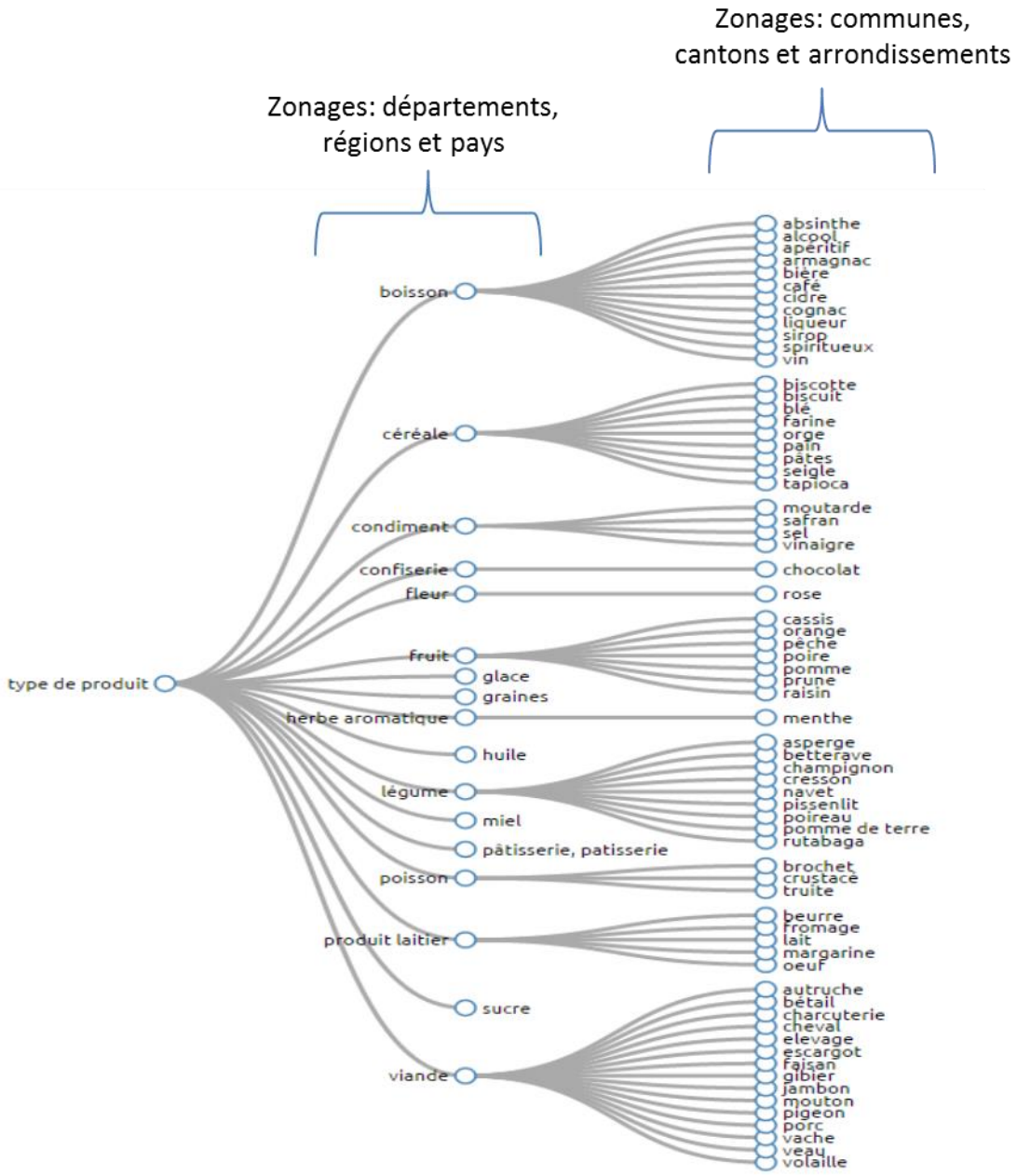


Figure 4.35 Taxonomies des types des produits fournis par les sites du patrimoine gourmand d'Ile de France.

4.4 Conclusion du chapitre

Dans ce chapitre nous avons tout d'abord présenté une approche d'interconnexion de données géoréférencées qui s'appuie sur les métadonnées sur les caractéristiques des géométries pour adapter dynamiquement le paramétrage de la comparaison géométrique. Cette approche permet de résoudre les problèmes liés à l'hétérogénéité géométrique en s'appuyant sur l'utilisation d'une base de règles pour déduire, au niveau de chaque comparaison d'une paire de géométries, les paramètres d'interconnexion. Elle permet donc à l'expert de paramétrer l'interconnexion à un niveau plus fin que l'approche classique où les paramètres de comparaison sont fixés a priori. Nous avons montré, à travers l'exemple des monuments parisiens, l'amélioration des résultats d'interconnexion apportée par cette approche par rapport à l'approche classique. Bien que nous ayons optimisé l'implémentation de l'approche pour réduire sa complexité en temps, elle pourrait faire l'objet d'autres améliorations pour réduire cette complexité. L'approche doit également être testée avec des jeux de données plus volumineux, notamment pour évaluer le passage à l'échelle de son implémentation dans la version MapReduce de Silk. Cette approche est beaucoup plus adaptée pour découvrir des relations de cardinalités 1:1 entre jeux de données présentant une forte densité spatiale et une forte hétérogénéité géométrique interne. À travers l'exemple des monuments parisiens, nous avons constaté que notre approche, ainsi que l'approche classique, ne sont pas adaptées à un cas récurrent de l'interconnexion de sources de données hétérogènes qui consiste à découvrir des relations n:m. Nous avons donc proposé une deuxième approche, complémentaire à l'approche adaptative, qui permet de découvrir des relations de cardinalité n:m. Cette deuxième approche utilise les données topographiques de référence comme support pour l'interconnexion. Les ressources sont d'abord ancrées à des données topographiques de référence puis des relations sont dérivées entre les ressources ancrées aux mêmes objets topographiques. Nous avons appliqué cette deuxième approche sur le cas des monuments parisiens. Les résultats montrent la plus-value de cette approche : elle permet d'augmenter le rappel tout en minimisant la perte en précision par rapport à l'approche classique. Les approches d'interconnexion proposées ici montrent un grand potentiel pour l'interconnexion de ressources géoréférencées dans le cas d'hétérogénéités géométriques inter et intra sources de données. Les résultats de ces approches doivent toutefois être consolidés par d'autres cas d'application.

Nous avons ensuite proposé deux applications de visualisation cartographique multi-échelle des données liées sur le Web en s'appuyant sur des données géographiques de référence. La première application s'appuie sur les liens d'ancrage, entre des ressources du Web de données et des données géographiques de référence dotées de géométries détaillées. Ces géométries permettent de porter les informations thématiques issues des ressources du Web de données, et améliorent donc leur lisibilité. Nous avons montré, à travers l'exemple des monuments parisiens issus de la base Mérimée, comment ces géométries détaillées permettent de proposer une visualisation cartographique multi-échelle en utilisant des techniques de généralisation de l'état de l'art. La seconde application s'appuie sur l'utilisation d'un maillage territorial multi-échelle pour proposer une solution d'exploration cartographique de n'importe quelle source de données géoréférencées. Cette application permet, d'une manière interactive, d'explorer un point d'accès Sparql et de sélectionner les ressources et leurs propriétés à visualiser. Elle s'appuie sur les liens entre les ressources du Web de données et les données du maillage de référence pour proposer à l'utilisateur de sélectionner une des méthodes de visualisation adaptées à la nature de la propriété sélectionnée. Ces méthodes de visualisation permettent notamment d'agréger les données et proposer des cartes statistiques sur les

valeurs de la propriété sélectionnée. L'une des plus-values de cette approche est la prise en compte de la hiérarchie des valeurs des données qualitatives classifiées ou ordonnées. L'approche permet d'affecter chaque niveau de la hiérarchie des concepts utilisés comme valeurs de propréité à un niveau de maillage dans la visualisation. Les deux applications proposées ici ont pour objectif de montrer que la visualisation cartographique multi-échelle des ressources géoréférencées peut être améliorée en utilisant des données géographiques de référence et en exploitant mieux la structure des données à visualiser. Elles ont néanmoins besoin d'être améliorées et complétées, notamment l'application d'exploration cartographique où plus de méthodes d'agrégation et de techniques de symbolisation pourraient être implémentées.

Conclusion de la partie B

Cette partie a regroupé les différentes solutions que nous avons proposées pour remédier aux problèmes liés à l'utilisation des références spatiales des ressources sur le Web comme critère pour leur interconnexion ainsi que comme information de localisation mise en œuvre dans des applications de visualisation cartographique des données.

Afin de résoudre les problèmes liés aux différences de niveaux de détail et de modélisation géométriques des références spatiales d'une ressource à une autre, nous avons proposé un vocabulaire qui permet d'expliciter, au niveau de chaque référence spatiale, les métadonnées sur ses caractéristiques : précision planimétrique, modélisation géométrique et caractère plus ou moins vague de l'entité géographique représentée. Nous avons ensuite proposé une approche qui permet d'acquérir ces métadonnées pour chaque référence spatiale dans un jeu de données à partir de données topographiques de référence.

La formalisation du niveau de détail et de la modélisation géométrique nous à permis de prendre en compte ces caractéristiques pour résoudre les problèmes liés au paramétrage d'un processus d'interconnexion qui utilise la comparaison des géométries comme critère de mise en correspondance. Associer à chaque géométrie la description de ses caractéristiques nous permet d'adapter dynamiquement le paramétrage de la comparaison de chaque paire de références géométries dans un processus d'interconnexion. Cette approche a été implémentée sous Silk. Les tests réalisés sur deux jeux de données sur les monuments parisiens ont permis de montrer une amélioration significative des résultats dans le cas d'une recherche de liens de cardinalité 1:1 entre des jeux de données présentant une forte hétérogénéité interne au niveau des géométries et une forte densité spatiale.

Nous avons ensuite proposé une approche d'interconnexion qui s'appuie sur des données topographique de références pour découvrir des relations de cardinalité n:m. Cette approche permet d'améliorer les résultats d'interconnexion par la découverte de nouveaux liens. Elle permet d'améliorer le rappelle tout en réduisant la perte en précision par rapport à l'approche classique. Les caractéristiques formalisées des géométries sont utilisée d'une manière implicite dans cette approche pour permet de choisir les des données topographiques de référence adéquates à utiliser comme source de support pour l'interconnexion.

Les solutions de visualisation cartographique multi-échelle que nous avons proposées dans cette partie s'appuient sur des liens créés entre données thématiques et données topographiques de référence. Dans la première application, le niveau de détail des géométries de données topographiques de référence permet de proposer à l'utilisateur une visualisation cartographique, de l'information thématique associée, d'une manière lisible et conviviale à différentes échelles. L'utilisation des données de maillage territorial nous permet de proposer une application d'exploration cartographique des données thématiques générique au sein de laquelle l'information thématique que l'utilisateur choisit de visualiser est agrégée à la volée pour proposer des cartes statistiques à différentes échelles. L'application offre à l'utilisateur le moyen de choisir la technique d'agrégation adaptée qui lui permet comprendre les données et leur distribution dans le maillage.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Rappel des objectifs

Dans cette thèse nous nous sommes intéressés aux données géoréférencées sur le Web de données, plus particulièrement à l'utilisation des références spatiales comme critère de décision pour la détection automatique de relations de correspondance pour l'interconnexion de données et comme information de localisation pour la visualisation cartographique des ressources géoréférencées. Nous avons vu que les questions liées au niveau de détail et à la modélisation géométrique des références spatiales constituent des défis pour l'interconnexion comme pour la visualisation.

En effet, la modélisation géométrique et le niveau de détail sont des caractéristiques des géométries utilisées pour géoréférencer les ressources qui peuvent être hétérogènes d'une source de données à une autre. Ce problème est connu dans le domaine de l'appariement des données géographiques. De plus, la nature ouverte des sources de données du Web, souvent d'origines diverses, rajoute un défi particulier à l'interconnexion : les caractéristiques des références spatiales peuvent être hétérogènes à l'intérieur d'une même source, ce qui rend le paramétrage des algorithmes de détection de relations de correspondance très complexe, voire impossible. L'objectif principal visé dans cette thèse est donc de proposer une approche d'interconnexion dont le paramétrage puisse être adapté finement pour surmonter les difficultés de détection de relations de correspondance dues aux hétérogénéités des références spatiales à traiter.

La visualisation cartographique des ressources dotées de références spatiales est une pratique courante qui permet d'explorer et de mieux appréhender les données. Dans la plupart des sources de données, les références spatiales sont modélisées par des points. Le niveau de détail de ces derniers ne permet de proposer une visualisation cartographique qui reste lisible qu'à grande échelle. Un second objectif visé dans cette thèse est de proposer des solutions de visualisation cartographique capables d'adapter automatiquement de niveau de détail de l'information affichée lors du changement de l'échelle de visualisation.

À travers les solutions proposées dans cette thèse, nous nous sommes attachés à la réalisation de ces objectifs.

Contributions

- **Un modèle pour la représentation des caractéristiques des géométries** : La première étape de notre travail a consisté à identifier les caractéristiques des géométries et à les formaliser au sein d'un vocabulaire de la sémantique des XY. Nous nous sommes appuyés sur les standards et les travaux existants pour la représentation de la qualité, du niveau de détail et de la modélisation géométrique des données géographique. La particularité de ce vocabulaire est de permettre d'associer à chaque référence spatiale d'une ressource les métadonnées qui décrivent ses caractéristiques. Ceci permet d'appréhender le sens de chaque référence spatiale, et ainsi gérer le problème de leur hétérogénéité à l'intérieur d'une source de données.
- **Une approche pour l'acquisition des caractéristiques des géométries :** Nous avons répondu au besoin d'acquérir les connaissances utiles sur les caractéristiques des géométries grâce à

une approche qui permet d'associer chaque géométrie d'un jeu de données à un type de modélisation géométrique et de déduire sa précision planimétrique. Cette approche s'appuie sur des données topographiques de référence qui décrivent, d'une manière détaillée, le contexte géographique des ressources géoréférencées. L'acquisition des différentes modélisations géométriques des références spatiales par la méthode proposée ici est effectuée grâce aux algorithmes de classification supervisée de la littérature, par apprentissage sur les relations spatiales entre ces références spatiales et les géométries des données topographiques de référence.

- **Une adaptation dynamique à un niveau fin des paramètres de mise en correspondance de ressources dotées de références spatiales:** Notre troisième contribution est la proposition d'une approche qui permet de résoudre le problème du paramétrage de la comparaison des références spatiales dans un processus d'interconnexion, lorsque celles-ci présentent une forte hétérogénéité au sein de leur source d'origine et une forte densité spatiale. Cette approche permet d'adapter dynamiquement, au niveau de chaque comparaison d'une paire de références spatiales, les paramètres de calcul de la similarité entre elles. Cette adaptation est réalisée grâce à une base de règles qui s'appuie sur les caractéristiques des deux références spatiales comparées pour inférer les paramètres de calcul de la similarité. Cette base de règles doit être définie par un expert en interconnexion. Cette approche permet d'expliciter les connaissances de l'expert en interconnexion et de les appliquer à un niveau très fin. Elle permet donc de surmonter le défi majeur de définir les paramètres d'interconnexion dans le cas d'une forte hétérogénéité des géométries à l'intérieur, ainsi qu'entre les jeux de données. Nous avons veillé à ce que l'implémentation de cette approche soit réalisée dans un outil d'interconnexion générique afin qu'elle soit utilisable pour le plus possible de cas d'application.
- **Une approche pour la détection des relations de cardinalité n:m entre les ressources dotées de références spatiales:** Notre quatrième contribution est la proposition d'une approche qui permet de découvrir des relations de cardinalité n:m dans un processus d'interconnexion de ressources grâce à leurs références spatiales. Cette approche utilise les données topographiques de référence comme ressources de support pour l'interconnexion. Dans cette approche, les données à interconnecter sont tout d'abord ancrées aux données topographiques de référence. Nous dérivons ensuite des liens entre les ressources lorsqu'elles sont ancrées aux mêmes ressources topographiques de référence. Le choix des données topographiques de référence à utiliser dans cette approche se fait grâce à des connaissances sur nature des ressources à interconnecter et aux métadonnées sur les caractéristiques de leurs références spatiales.
- **Exploitation de liens entre ressources thématiques et ressources topographiques pour la visualisation cartographique multi-échelle :** Nous avons proposé une première application de visualisation cartographique interactive qui offre une meilleure lisibilité des informations portées par les ressources thématiques. Cette approche s'appuie sur les liens créés entre ressources thématiques et ressources topographiques afin de combiner les informations thématiques des premières avec les géométries détaillées des dernières. Le niveau de détail élevé des géométries des données topographiques permet dans cette application de fournir des cartes thématiques lisibles à grande échelle et l'utilisation de techniques de généralisation de l'état de l'art sur ces références spatiales détaillées permet de fournir de nouvelles géométries, simplifiées, lisibles à plus petite échelle.

- **Exploration cartographique des données thématique grâce à des données géographiques de maillage :** Nous avons proposé une seconde application de visualisation cartographique multi-échelle qui permet d'explorer les différentes sources de données géoréférencées sur le Web. Elle s'appuie sur des données topographiques de maillage afin d'agréger l'information thématique portée par les ressources de la source explorée. Cette application permet à l'utilisateur de spécifier le point d'accès de la source de données et d'explorer les types de ressource ainsi que les propriétés qui les décrivent. L'application offre à l'utilisateur le choix de la technique d'agrégation adaptée à la nature de la propriété qui porte l'information thématique à visualiser. L'utilisation des données de maillage permet de simplifier l'information thématique à afficher tout en fournissant un niveau de détail géométrique adapté à chaque échelle de visualisation. L'une des originalités de cette application est la prise en compte de la sémantique des valeurs des propriétés: l'application offre le choix d'un mode dynamique d'agrégation qui permet d'associer chaque niveau du maillage à un niveau dans la taxonomie qui structure les valeurs de la propriété thématique à agréger.

Dans ce travail, nous soulignons l'importance des données topographiques de référence à travers leur utilisation répétée. En effet, le niveau de détail géométrique et le niveau d'exhaustivité garantis par les données topographiques de référence dans leur description de l'espace géographique font de ces données des sources de connaissances de qualité suffisante pour nous permettre de proposer les différentes approches présentées dans cette thèse.

Perspectives

Les solutions que nous avons proposées dans ce travail ont permis de répondre à nos objectifs. Plusieurs améliorations peuvent toutefois être apportées à ces solutions. En outre, d'autres applications dans des contextes différents que celui présenté dans cette thèse peuvent être envisagées pour les différentes propositions de ce travail.

L'approche d'acquisition des métadonnées de modélisation géométrique de chaque géométrie dans une source de données a été mise en œuvre dans cette thèse selon plusieurs étapes en utilisant des outils différents. Une amélioration possible de cette approche est de proposer un seul outil qui automatise la globalité de son processus. Cette automatisation doit prendre en compte la question du passage à l'échelle, car les différentes opérations de calcul des descripteurs et des modèles d'apprentissage peuvent être très coûteuses en temps et en mémoire. De plus, nous avons vu que cette approche s'appuie sur des hypothèses sur les différents choix de modélisation géométrique faits par les contributeurs qui ont saisi les données. Ces choix constituent les classes possibles de la classification supervisée. Une perspective qui serait pertinente à explorer est d'étudier la possibilité de proposer une approche sans hypothèses sur les choix de modélisation en s'appuyant sur des algorithmes de classification non supervisée. Une perspective qui peut être intéressante pour cette approche est de l'appliquer dans un autre contexte. En effet, associer à chaque géométrie la description de ses caractéristiques de modélisation géométrique et de précision planimétrique dans une source de données peut être utilisé dans des buts autres que l'interconnexion. Par exemple, cela peut être utilisé pour estimer le niveau de cohérence des données dans une source de données, ou pour qualifier les données de plateformes participatives comme OpenStreetMap pour harmoniser leur contenu.

Pour mieux démontrer l'efficacité de l'approche d'adaptation dynamique des paramètres de comparaison pour chaque paire de géométries, celle-ci doit être utilisée pour d'autres cas d'application, notamment dans le cadre d'une interconnexion multicritère. Cette approche reste ouverte à d'autres améliorations également. Par exemple, nous fournissons un moyen à l'expert de en interconnexion de définir le comportement de la fonction de similarité en adaptant les paramètres de convexité et sa concavité ainsi que le seuil de distance selon les caractéristiques des géométries comparées. Ces paramètres sont ensuite utilisés dans la définition de la fonction de similarité. Pour plus de flexibilité, une amélioration possible de cette approche serait de laisser la main à l'utilisateur pour définir, à travers les règles, toute l'équation de la fonction de similarité. L'implémentation de cette approche reste ouverte à des améliorations pour réduire sa complexité temporelle. Nous pouvons envisager l'utilisation d'un cache de résultats de raisonnement sur les règles afin réduire le nombre d'appels du moteur de raisonnement durant la phase de mise en correspondance. En outre, cette implémentation a été adaptée à la version MapReduce de Silk. Le passage à l'échelle de cette implémentation doit également être vérifié en évaluant ses résultats sur des jeux de données volumineux.

La méthode d'adaptation du paramétrage de la comparaison de valeurs de propriétés à niveau fin est appliquée dans ce travail pour le cas des géométries, mais peut être envisagée pour d'autres critères d'interconnexion. En effet, s'appuyer sur les métadonnées qui décrivent la qualité et les conditions d'acquisition des valeurs des propriétés afin d'adapter leur comparaison est une idée qui peut être employée sur d'autres propriétés quand les valeurs de celles-ci sont représentées d'une manière hétérogène à l'intérieur d'une même source. Par exemple, l'application de cette approche pourrait être utile dans le cas où nous voudrions utiliser la comparaison de propriétés qui décrivent des dates pour interconnecter les ressources de deux jeux de données différents, en sachant que les dates sont formatées d'une manière hétérogène à l'intérieur d'un des jeux de données. Dans ce cas, disposer d'une métadonnée qui spécifie le format de la date pour chaque valeur de date permet de le prendre en compte lors de la comparaison en homogénéisant le format des deux dates avant de calculer la distance entre elles.

Des améliorations peuvent être apportées également à l'approche d'interconnexion par ancrage-dérivation que nous avons proposée pour la découverte de relations de cardinalité n:m. Une automatisation globale de l'approche peut être envisagée pour faciliter son application. On peut imaginer une approche qui permet de choisir automatiquement les types des données topographiques à utiliser comme support pour l'interconnexion. L'automatisation de l'enchaînement des différentes étapes d'ancrage et de dérivation peut également être envisagée. Dans le cas de la mise en œuvre de cette approche d'interconnexion, nous avons utilisé des liens de type owl:sameAs pour des raisons pratiques. Une perspective importante pour cette approche serait de définir des liens qui portent une sémantique beaucoup plus appropriée entre les données thématiques et les données topographiques de référence.

Les prototypes d'applications de visualisation cartographiques présentés dans cette thèse, que nous utilisions localement, doivent être complétés et mis en ligne afin tester leur robustesse. De plus, plusieurs améliorations peuvent être apportées aux approches de visualisation cartographique proposées. Les métadonnées qui décrivent le niveau de détail et la modélisation géométrique des références spatiales pourraient être mieux exploitées dans la génération des cartes. Elles permettraient, par exemple, d'automatiser les choix de l'échelle de visualisation à laquelle la

géométrie d'une ressource peut être affichée. L'utilisation des métadonnées de niveau de détail et de modélisation géométrique pour gérer l'hétérogénéité des géométries dans un processus de généralisation doit également être investiguée. L'une des perspectives les plus importantes pour l'application d'exploration cartographique produisant des cartes statistiques est de mieux exploiter la sémantique de l'information thématique visualisée. Il faut donc mieux étudier les relations entre le changement de niveau de détail du maillage territorial et le changement de niveau hiérarchique des nœuds dans la taxonomie qui structure les valeurs possibles de la propriété thématique choisie pour la visualisation.

LISTE DES FIGURES

Figure 1.1 La pile des technologies du Web sémantique (Semantic Web layercake diagram « https://www.w3.org/2007/03/layerCake.png ») **17**
Figure 1.2 Exemple de triplets dans un graphe RDF : en rouge le sujet, en jaune les prédicats et en vert les objets. **17**
Figure 1.3 Diagramme du nuage de données ouvertes liées (*Linking Open Data cloud diagram)* 2017 (Andrejs et al., 2017) **19**
Figure 1.4 Extrait de la visualisation de la ressource http://dbpedia.org/resource/Paris dans Lodlive **21**
Figure 1.5. Exemple de représentations littérale et structurée de la géométrie de la ressource qui décrit le département du Val-de-Marne http://data.ign.fr/id/geofla/departement/94 ..... **23**
Figure 1.6 Exemples d'hétérogénéités géométriques des données géographiques (Abadie, 2012) **33**
Figure 1.7 Exemple d'une imprécision de localisation des ressources géoréférencées du Web de données : la ressource DBpedia décrivant le Musée de l'Ordre de la Libération. **34**
Figure 1.8 Exemple de différence de modélisations géométriques des références spatiales des ressources qui décrivent Londres entre DBpedia et l'Ordnance Survey. **35**
Figure 1.9 Exemple de différence de granularités géométriques entre les sources de données GDAM et NUTS **35**
Figure 1.10 Exemple de différence de localisation d'un col de montagne entre la BD TOPO® et la BD CARTO® à cause du caractère vague de ce type d'entités géographiques. **36**
Figure 1.11 Exemple d'hétérogénéité géométrique à l'intérieur d'une même source de données **37**
Figure 1.12 Exemple de solutions de visualisation cartographique propres à des sources de données géoréférencées du Web de données. **38**
Figure 1.13 Visualisation des ressources YAGO qui représentent les entreprises fondées dans la région de la baie de San Francisco durant le dernier siècle (Hoffart et al., 2013). **39**
Figure 1.14 Exemple de solutions de visualisation cartographique qui permettent l'exploration des sources de données géoréférencées quelconques du Web de données **40**
Figure 2.1 Exemples d'opérateurs de généralisation issus de la classification de (Mustière, 2001) **45**
Figure 2.2 Exemple explicatif de la distance de *Hausdorff* **46**
Figure 2.3 Exemple explicatif de la distance moyenne **46**
Figure 2.4 Exemple de cas où la distance de *Hausdorff* n'est pas adaptée pour comparer les géométries linéaires. On lui préfère alors la distance de *Fréchet*. **47**
Figure 2.5 Schéma général d'un processus d'interconnexion adapté du schéma proposé par (Ferrara et al., 2013) **55**
Figure 3.1 Exemple de ressource ayant plusieurs géométries : la ressource http://data.ign.fr/id/geofla/departement/75 représentant le département de Paris dans le jeu de données Geofla, la première géométrie représente la limite du département, la deuxième représente la localisation du siège de son chef-lieu (visualisation réalisées avec http://en.lodlive.it) **73**
Figure 3.2 Précision planimétrique des données BD TOPO® et la BD ADRESSE® selon leur source. Extrait des spécifications de la base BD TOPO® 2.2 (http://professionnels.ign.fr/doc/DC-BDTOPO-2-2.pdf) et de la BD ADRESSE® 2.1 (http://professionnels.ign.fr/doc/DC_BDADRESSE_2-1.pdf). **74**
Figure 3.3 Précision planimétrique des données BD CARTO®. Extrait des spécifications de la base BD CARTO®3.2 (http://professionnels.ign.fr/sites/default/files/DC_BDCARTO_3-2.pdf). **75**
Figure 3.4 Précision planimétrique des données BD ADRESSE® selon leurs types de localisation. Extrait des spécifications de la base BD ADRESSE®2.1 (http://professionnels.ign.fr/doc/DC_BDADRESSE_2-1.pdf). **75**
Figure 3.5 Extrait du vocabulaire de la sémantique des XY pour représenter la précision planimétrique des géométries. **77**
Figure 3.6 Représentation en RDF de la précision planimétrique de la géométrie délimitant le département 94 dans la source Geofla ign:94 ..... **77**

Figure 3.7 Extrait des spécifications de la classe BATIMENT de la base de données BD TOPO® 1.2 78
Figure 3.8 Extrait des spécifications de la classe TRONCON_COURS_EAU de la base de données BD TOPO® 2.2 (http://professionnels.ign.fr/sites/default/files/DC_BDTOPO_2-2.pdf) 79
Figure 3.9 Extrait de la fiche de métadonnées descriptives des jeux de données d'adressage du Grand Lyon.79
Figure 3.10 Extrait du vocabulaire de la sémantique des XY pour représenter la modélisation géométrique. 81
Figure 3.11 Représentation de la modélisation géométrique de la géométrie délimitant le département 94 dans la source Geofla http://data.ign.fr/id/geofla/departement/94 81
Figure 3.12 Extrait du vocabulaire de la sémantique des XY pour représenter le caractère vague des éléments caractéristiques de la forme de certains types d'entités géographiques. 82
Figure 3.13 Résolution géométrique représentée par une échelle équivalente. Extrait des métadonnées de la BD TOPO®. (http://professionnels.ign.fr/sites/default/files/métadonnées de produit BD TOPO®.html) 83
Figure 3.14 Extrait du vocabulaire de la sémantique des XY pour représenter la résolution géométrique. 83
Figure 3.15 Exemple de représentation en RDF de la résolution géométrique de la géométrie délimitant le département 94 dans la source Geofla® http://data.ign.fr/id/geofla/departement/94 84
Figure 3.16 Extrait du descriptif de contenu de la BD ADRESSE®2.2 pour les résultats du contrôle de qualité géométrique. 85
Figure 3.17 Exemple de positionnement des points adresse dans la BD ADRESSE® : par projection à partir du centroïde de la parcelle 86
Figure 3.18 Co-visualisation des localisations des monuments historiques de DBpedia (en jaune) avec un fond orthophotographique IGN 90
Figure 3.19 Extrait des géométries des adresses du grand Lyon à co-visualisées avec des données géographiques de référence de l'IGN 92
Figure 3.20 Descripteurs choisis pour l'apprentissage du modèle de classification des modélisations géométriques des adresses du Grand Lyon 93
Figure 3.21 Descripteurs spatiaux pour l'apprentissage du modèle de classification des modélisations géométriques des géométries des monuments DBpedia. 95
Figure 3.22 Fréquence des valeurs de précision planimétrique pour les monuments historiques de Paris de DBpedia en fonction du type de choix de modélisation. L'axe des abscisses indique les valeurs de précisions maximales, en mètres et celui des ordonnées le nombre de ressources DBpedia. 98
Figure 4.1 Calcul de similarité à partir de la valeur de distance spatiale dans LIMES 102
Figure 4.2 Calcul de similarité (confiance) dans Silk à partir de la valeur de distance d et du seuil Θ. 103
Figure 4.3 L'approche d'adaptation dynamique du paramétrage des comparaisons des références spatiales en fonction de leurs caractéristiques 106
Figure 4.4 Variation du comportement de la fonction de similarité 107
Figure 4.5 Étapes générales et classes principales dans la version SingleMachine de Silk 2.6 108
Figure 4.6 Exemple de paramétrage de mesure de distance numérique dans l'interface graphique de Silk . 109
Figure 4.7 Implémentation de l'approche d'adaptation dynamique des paramètres d'interconnexion 110
Figure 4.8 Exemple de l'usage de la mesure de distance géométrique rajoutée à Silk. Le code du système de coordonnées de référence ainsi que les coordonnées maximum et minimum du rectangle englobant sont fournis. 113
Figure 4.9 Exemple de l'utilisation de l'opérateur de transformation de coordonnées géographiques. Dans cet exemple, les coordonnées seront transformées du WGS84 vers Lambert93 113
Figure 4.10 Exemple d'une ressource Wikidata décrivant un monument historique parisien (l'Arc de Triomphe). Sa description comprend une référence vers l'identifiant du monument dans la base Mérimée. 116
Figure 4.11 Interface de visualisation de l'ensemble de liens de référence. Les points rouges représentent les monuments DBpedia, les verts représentent les monuments Mérimée. Les lignes jaunes représentent les liens. En cliquant sur un point, une fenêtre récupère à la volée et affiche quelques informations sur le monument concerné. 116

Figure 4.12 Résultats de l'approche classique (sans adaptation dynamique de paramètres) en termes de précision, de rappel et de F-mesure selon les valeurs de seuil. La valeur maximale de la F-mesure sera utilisée comme repère pour les résultats notre approche. 118
Figure 4.13 Résultats d'interconnexion par l'approche d'adaptation dynamique des paramètres de comparaison géométrique. 120
Figure 4.14 Résultats d'interconnexion par l'approche d'adaptation dynamique des paramètres de comparaison géométrique après correction des résultats d'acquisition automatique des caractéristiques des géométries. 121
Figure 4.15 Lien découvert seulement par l'approche adaptative. Les deux monuments représentent « l'Hôtel de la Marine » dans DBpedia (point rouge) et Mérimée (point vert). 123
Figure 4.16 Faux lien (entre deux monuments distincts) découvert par l'approche classique, mais évité par l'approche adaptative. Le point rouge représente le monument DBpedia. Le point vert représente le monument Mérimée. 123
Figure 4.17 Exemple de l'utilisation de données topographique de référence comme support pour l'interconnexion de ressources géoréférencées dans le Web 125
Figure 4.18 Etapes effectuées dans le cadre de l'approche d'interconnexion par ancrage-dérivation basée sur les données topographiques de référence. 127
Figure 4.19 Résultats d'interconnexion globaux obtenus par nos approches par rapport aux résultats obtenus par l'approche classique 128
Figure 4.20 Exemple de liens de cardinalité n:m découvert par notre approche d'ancrage-dérivation. 129
Figure 4.21 Exemple de nouveau lien découvert entre deux ressources très distantes spatialement grâce à notre approche d'ancrage-dérivation. 129
Figure 4.22 Exemple de faux lien découvert par l'approche classique, mais évité par notre approche d'ancrage-dérivation. Ceci est dû au fait d'ancrer des monuments proches mais distincts à des bâtiments différents. 130
Figure 4.23 Architecture de l'application de visualisation cartographique des données thématique grâce à leur interconnexion avec des données topographiques de référence. 132
Figure 4.24 Co-visualisation des données thématiques des deux sources Mérimée et DBpedia à grande échelle grâce aux liens d'ancrage avec les bâtiments des données topographiques de support BD PARCELLAIRE® 132
Figure 4.25 Résultats de l'application de l'algorithme de regroupement-amalgamation des géométries des bâtiments pour la visualisation des données thématiques Mérimée (siècles de construction des monuments) à l'échelle d'affichage des « quartiers » (entre les niveaux « rue » et « ville »). 134
Figure 4.26 Résultats de l'application de l'algorithme de regroupement-amalgamation des géométries des bâtiments pour la visualisation des données thématiques Mérimée (siècles de construction des monuments) à l'échelle d'affichage de la « ville » ou plus petite. 134
Figure 4.27 Solution de visualisation utilisant des techniques de sélection. À gauche une carte représentant la population par communes sur le site Géoclip. À droite une carte de localisation des sites touristique sur le site Tripadvisor. 136
Figure 4.28 Cartes répertoriant les offres immobilières en France sur le site Homengo : exemple d'agrégation par comptage. 136
Figure 4.29 Arbre de décision sur les possibilités de cartographie des données selon leur nature. 137
Figure 4.30 Aperçus et légende QGIS du jeu de données du « patrimoine gourmand » avec zoom sur Paris. Les couleurs représentent les différentes natures des sites représentés dans cette base. 139
Figure 4.31 Architecture 3-tiers de l'application d'exploration cartographique multi-échelle 139
Figure 4.32 Formulaire d'interaction avec l'utilisateur pour le choix de l'information thématique, sa nature et la technique d'agrégation multi-échelle. 141
Figure 4.33 Utilisation de la technique d'agrégation par comptage de valeurs distinctes d'une propriété représentées dans des cercles proportionnels pour la visualisation multi-échelle de données thématiques. La propriété sélectionnée dans cet exemple est l'identifiant « ref » des ressources. 143

Figure 4.34 Utilisation de la technique d'agrégation par comptage des valeurs dominantes représentées par des diagrammes circulaires de tailles proportionnées pour la visualisation multi-échelle de données. La propriété sélectionnée dans ce cas est le type du site du patrimoine gourmand. 144
Figure 4.35 Taxonomies des types des produits fournis par les sites du patrimoine gourmand d'Ile de France. 145

LISTE DES TABLEAUX

Tableau 3.1 Évaluation des résultats de la classification par apprentissage supervisé des géométries des adresses du Grand Lyon. 94
Tableau 3.2 Résultats fournis par Weka pour les quatre algorithmes de classification testés 96
Tableau 4.1 Évaluation des résultats de l'approche classique (sans adaptation dynamique de paramètres) en termes de précision, de rappel et de F-mesure selon les valeurs de seuil..... 117
Tableau 4.2 Évaluation des résultats d'interconnexion par l'approche d'adaptation dynamique des paramètres de comparaison géométrique. 120
Tableau 4.3 Évaluation des résultats d'interconnexion par l'approche d'adaptation dynamique des paramètres de comparaison géométrique après correction des résultats d'acquisition automatique des caractéristiques des géométries..... 121
Tableau 4.4 Évaluation des résultats d'interconnexion globaux obtenus par nos approches par rapport aux résultats obtenus par l'approche classique. 128
Tableau 4.5 Association entre niveaux de zoom de visualisation et niveaux de données de zonage. 140

BIBLIOGRAPHIE

Abadie, N. Formalisation, acquisition et mise en œuvre de connaissances pour l'intégration virtuelle de bases de données géographiques: les spécifications au cœur du processus d'intégration (Doctoral dissertation, Université Paris-Est), **2012**

Abu Helou, M. & Palmonari, M. Upper bound for cross-lingual concept mapping with external translation resources. International Conference on Applications of Natural Language to Information Systems. Springer, Cham, **2015**

Achichi, M., Bellahsene, Z. & Todorov, K. A survey on web data linking. Revue des Sciences et Technologies de l'Information-Série ISI: Ingénierie des Systèmes d'Information, **2016**

Ahlers, D. Assessment of the accuracy of geonames gazetteer data. In Proceedings of the 7th Workshop on Geographic Information Retrieval. ACM, **2013,** pp. 74–81

Aleksovski, Z. Ten Kate, W. & Van Harmelen, F. Exploiting the structure of background knowledge used. Proceedings of the 1st International Conference on Ontology Matching. CEUR-WS. org, **2006**, Vol.225

Alt, H. & Godau, M. Computing the Fréchet distance between two polygonal curves. International Journal of Computational Geometry and Applications, **1995**, Vol. 5(1-2), pp. 75-91

Andrejs, A., McCrae J. P., Buitelaar, P. Jentzsch, A. & Cyganiak, R. Linking Open Data cloud diagram **2017**. Disponible sur: http://lod-cloud.net/

Araujo, S., Vries, A. D. & Schwabe, D. (2011, October). Serimi results for OAEI 2011. In Proceedings of the 6th International Conference on Ontology Matching CEUR-WS. Org, **2011**, Vol. 814, pp. 212-219

Arkin, E.M., Chew, L.P., Huttenlocher, D.P., Kedem, K., Kedem, K. & Mitchel, J.S.B. An efficient computable metric for comparing polygonal shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, **1991**, Vol. 13(3), pp. 209-216

Atemezing, G. A. & Troncy, R. Comparing vocabularies for representing geographical features and their geometry. Terra Cognita 2012 Workshop, **2012**, Vol. 3

Atencia, M., David, J. & Scharffe, F. Keys and Pseudo-Keys Detection for Web Datasets Cleansing and Interlinking. In EKAW, 2012(10), pp. 144-153

Auer, S., Lehmann, J. & Hellmann, S. Linkedgeodata: Adding a spatial dimension to the web of data. The Semantic Web-ISWC 2009, **2009**, pp. 731-746

Barron, C., Neis, P. & Zipf, A. A comprehensive framework for intrinsic openstreetmap quality analysis. Transactions in GIS, **2014**, Vol. 18(6), pp. 877–895.

Bel Hadj Ali, A. Qualité géométrique des entités géographiques surfaciques - Application à l'appariement et définition d'une typologie des écarts géométriques. Thèse de Doctorat. Université Marne-la-Vallée, **2001**

Berners-Lee, T. J. Information management: A proposal. No. CERN-DD-89-001-OC. **1989**

Berners-Lee, T. J. Linked data-design issues. **2006**. Disponible sur: http://www. w3. org/DesignIssues/LinkedData. Html

Bizer, C, Heath, T. & Berners-Lee, T. Linked data-the story so far. Semantic services, interoperability and web applications: emerging concepts, **2009**, pp. 205-227

Bouquet, P., Stoermer, H. & Bazzanella, B. An Entity Naming System for the Semantic Web. In Proceedings of the 5th European Semantic Web Conference (ESWC 2008), **2008**

Brando, C. Un modèle d'opérations réconciliables pour l'acquisition distribuée de données géographiques. Thèse de doctorat de l'université Paris Est, **2013**

Buscaldi, D. Approaches to disambiguating toponyms. Sigspatial Special, 2011, Vol. 3(2) pp. 16-19.

Campedel, M. Classification supervisée. Telecom Paris, **2005**

Chen, S., Ma, B. & Zhang, K. On the similarity metric and the distance metric. Theoretical Computer Science, Elsevier, **2009**, pp. 410, 2365-2376

Christen, P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media, **2012**

Cohen, S.D. & Guibas, L.J. Partial matching of planar polylines under similarity transformations. In Proceeding of the 8th Annual ACMSIAM Symposium on Discrete Algorithms, **1997**, pp. 777-786

Cohen, W., Ravikumar, P. & Fienberg, S. A comparison of string metrics for matching names and records. In Kdd workshop on data cleaning and object consolidation, **2003**(8), Vol. 3, pp. 73-78

Costes, B. Matching old hydrographic vector data from Cassini's maps , e-Perimetron, **2014**, Vol. 9(2), pp. 51—65

de León, A., Saquicela, V., Vilches, L. M., Villazón-Terrazas, B., Priyatna, F. & Corcho, O. Geographical linked data: a Spanish use case. In Proceedings of the 6th International Conference on Semantic Systems. ACM, **2010**, pp. 36

Deakin, R. E. & Hunter, M. N. Geodesics on an ellipsoid – Pittman's method, Proceedings of the Spatial Sciences Institute Biennial International Conference (SSC2007), Hobart, Tasmania, Australia, **2007**, pp. 223-242

Debattista, J., Lange, C. & Auer, S. daQ, an Ontology for Dataset Quality Information. In LDOW, 2014

Dempster, A. P. A generalization of bayesian inference. Journal of theRoyal Statistical Society. Series B (Methodological), **1968**, pp. 205–247.

Devogele, T. Processus d'intégration et d'appariement des bases de données géographiques - Application à une base de données routière multi-échelles. Thèse de doctorat. Université de Versailles, Laboratoire COGIT, Institut Géographique National, **1997**.

Devogele T., Parent C. and Spaccapietra S. On spatial Database Integration. International Journal of Geographic Information Science - Special Issue: Interoperability in GIS, Taylor & Francis, Vol. 12(4) **1998**. pp. 335-35.

Duchateau, F., Bellahsene, Z. & Roche, M. A Context-based Measure for Discovering Approximate Semantic Matching between Schema. In RCIS'07: Research Challenges in Information Science, **2007**

Duchateau, F., Bellahsene, Z. & Coletta, R. A flexible approach for planning schema matching algorithms. On the Move to Meaningful Internet Systems: OTM (2008), **2008**, pp. 249-264

Duckham, M., Kulik, L., Worboys, M. & Galton, A. Efficient generation of simple polygons for characterizing the shape of a set of points in the plane. Pattern Recognition, **2008**, Vol. 41(10), pp. 3224-3236

Dumenieu, B. Un système d'information géographique pour le suivi d'objets historiques urbains à travers l'espace et le temps (Doctoral dissertation, Paris, EHESS), **2015**

Egenhofer, M. & AlTaha, K. Reasoning about Gradual Changes of Topological Relationships in Theory and Methods of Spatio-Temporal Reasoning in Geographic Space, Pisa, Italy, and in: Frank A., Campari I., and Formentini U. (eds.), Lecture Notes in Computer Science, Springer-Verlag, **1992,** Vol**.** 639, pp. 196-219

Egenhofer, M. J. & Franzosa, R. Point-set topological spatial relations. Int. Journal of Geographical Information Systems 5(2), **1991**, Vol. 5(2), pp.161-174

Egenhofer, M. J. & Herring, J. R. Categorizing binary topological relations between regions, lines, and points in geographic databases. **1990**, The, Vol.9(94-1), pp. 76

Egenhofer, M. J. & Herring, J. R. Categorizing binary topological relationships between regions, lines and points in geographic databases. Tech. Report, Department of Surveying Engineering, **1991**, pp. 1-28

Elmagarmid, A. K., Ipeirotis, P. G. & Verykios, V. S. Duplicate record detection: A survey. IEEE Transactions on knowledge and data engineering, **2007**, Vol. 19(1), pp. 1-16.

Euzenat, J., & Shvaiko, P. Ontology matching. Heidelberg: Springer, **2007**, Vol. 18.

Fan, Z., Euzenat, J. & Scharffe, F. Learning concise pattern for interlinking with extended version space. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE/WIC/ACM International Joint Conferences on. IEEE, **2014**, Vol.1

Fellbaum, C. WordNet. John Wiley & Sons, Inc., **1998**

Ferrara, A., Nikolov, A. & Scharffe, F. Data linking for the semantic web. Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications, **2013**, pp. 326

Fonseca, F., Davis, C., & Câmara, G. Bridging ontologies and conceptual schemas in geographic information integration. GeoInformatica, **2003**, Vol. 7(4), pp. 355-378.

Gandon, F., Corby, O. & Faron-Zucker, C. Le web sémantique : Comment lier les données et les schémas sur le web?. Dunod, **2012**

Georgala, K. Proceedings of the The 15th International Semantic Web Conference (ISWC2016), Doctoral Consortium Track, Kobe, Japan, **2016**

George-Nektarios, T. Weka classifiers summary. Athens University of Economics and Bussiness Intracom-Telecom, Athens, **2013**

Gesbert, N. Etude de la formalisation des spécifications de bases de données géographiques en vue de leur intégration. Thèse de doctorat. Université de Marne-la-Vallée, **2005**

Gill, L. OX-LINK: The Oxford medical record linkage system, **1997**

Girres, J. F. & Touya, G. Quality assessment of the french openstreetmap dataset. Transactions in GIS, **2010**, Vol. 14(4), pp. 435–459

Girres, J. F. Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques. Doctoral dissertation, Paris Est, **2012**

Goldberg, D. W., Wilson, J. P. & Knoblock, C. A. From text to geographic coordinates: the current state of geocoding. URISA journal, **2007**, Vol. 19(1), pp. 33-46

Grabisch, M. & Perny, P. Agrégation multicritère. Logique floue, principes, aide à la décision, **2003**, pp. 81-120

Gruber, T. R. Toward principles for the design of ontologies used for knowledge sharing?. International journal of human-computer studies, **1995**, Vol. 43(5-6), pp. 907-928

Hahmann, S. & Burghardt, D. Connecting linkedgeodata and geonames in the spatial semantic web. In 6th International GIScience Conference, **2010**

Hamdi, F., Abadie, N., Bucher, B. & Feliachi, A. Geomrdf: A geodata converter with a fine-grained structured representation of geometry in the web. arXiv preprint arXiv:1503.04864, **2015**

Hamming, R. Error detecting and error correcting codes, Technical Report 2. Bell System Technical Journal, **1950**

Hauff, C. A study on the accuracy of flickr's geotag data. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, **2013**, pp. 1037–1040

Hirst, G. & St-Onge, D. Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An electronic lexical database, **1998**, Vol. 305, pp. 305-332

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, Vol.194, **2013**, p. 28-61.

Hu, W., Chen, J. & Qu, Y. A self-training approach for resolving object coreference on the semantic web. In Proceedings of the 20th international conference on World wide web. ACM, **2011**, pp. 87-96

Huza, M., Harzallah, M. & Trichet, F. OntoMas: a tutoring system dedicated to ontology matching. Proceedings of the 1st International Conference on Ontology Matching. CEUR-WS. Org, **2006**, Vol. 225

Hyland, B., Atemezing, G. & Villazon-Terrazas, B. Eds. Best Practices for Publishing Linked Data, W3C Working Group Note. **2014**. Disponible sur: http://www.w3.org/TR/ld-bp/

Isele, R. & Bizer, C. Active learning of expressive linkage rules using genetic programming. Web Semantics: Science, Services and Agents on the World Wide Web, **2013**, Vol. 23, pp. 2-15

Isele, R. & Bizer, C. Learning linkage rules using genetic programming. Proceedings of the 6th International Conference on Ontology Matching. CEUR-WS. org, **2011**, Vol. 814

Isele, R., Jentzsch, A. & Bizer, C. Efficient Multidimensional Blocking for Link Discovery without losing Recall. In WebDB, **2011**

ISO 19101: Geographic information - Reference model. International Organization for Standardization (TC 211), **2002** (19101)

ISO 19107: Geographic information - Spatial Schema. International Organization for Standardization (TC 211), **2003** (19107)

ISO 19109: Geographic information - Rules for application schema. International Organization for Standardization (TC 211), **2005** (19109).

ISO 19112: Geographic information -- Spatial referencing by geographic identifiers. International Organization for Standardization (TC 211), **2003** (19112)

ISO 19115: COX, S. J. D., OWL representation of Geographic Information - Metadata - Data quality package. **2003**. Disponible sur: http://def.seegrid.csiro.au/isotc211/iso19115/2003/dataquality

ISO 19115: Geographic information - Metadata. International Organization for Standardization (TC 211), **2003** (19115)

ISO 19115-1: Geographic information - Metadata - Part 1: Fundamentals. International Organization for Standardization (TC 211), **2014** (19115-1)

ISO 19131: Geographic information - Draft international standard. International Organization for Standardization (TC 211), **2007** (19131)

ISO 19139: Geographic information - Metadata, Implementation specification. International Organization for Standardization (TC 211), **2007** (19139)

ISO 19157: Geographic information – data quality. International standard, International Organization for Standardization (TC 211), **2013**(19157)

ISO 8402: International Organization for Standardization. Quality Management and Quality Assurance-Vocabulary. **1994**(8402)

ISO 8402: Quality Management and Quality Assurance-Vocabulary. International Organization for Standardization. **1994** (8402)

Jaara, K. Prise en compte des dépendances entre données thématiques utilisateur et données topographiques lors d'un changement de niveau de détail (Doctoral dissertation, Paris Est), **2015**

Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat, **1901**, Vol. 37, pp. 547-579

Jaffri, A., Glaser, H., & Millard, I. Uri disambiguation in the context of linked data, **2008**

Jaro M.A. Unimatch: A Record Linkage System: User's Manual. technical report, US Bureau of the Census, Washington, D.C., **1976**

Kavouras, M., & Kokla, M. Semantic integration of heterogeneous geospatial information. Advances in Photogrammetry, Remote Sensing and Spatial Information Sciences, **2008**, pp. 303-310

Kondrak, G. N-gram similarity and distance. String processing and information retrieval. Springer Berlin/Heidelberg, **2005**

Lemarié C., and Raynal L. Geographic data matching : First investigations for a generic tool. In Proceedings of GIS/LIS 96, ACSM Annual conference and Exposition, Denver, **1996,** pp. 405-420

Lenzerini, M. Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, **2002**. pp. 233-246.

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, **1966**(2), Vol. 10(8), pp. 707-710

Li, L. & Goodchild, M. F. An optimisation model for linear feature matching in geographical data conflation. International Journal of Image and Data Fusion, **2011**, Vol. 2(4), pp. 309–328

Mackaness, W., Burghardt, D. & Duchêne, C. Map Generalisation: Fundamental to the modelling and understanding of geographic space. Chapter 1 in D. Burghardt, C. Duchene and W.A. Mackaness (eds): Abstracting Geographic Information in a Data Rich World, LNG&C, Springer, Heidelberg, **2003**, ISBN 978-3-319-00202-6

McMaster, R. A. Statistical Analysis of Mathematical Measures for Linear Simplification, The American Cartographer, **1986**, Vol. 23

Méneroux, Y., & Brasebin, M. Towards a generic method for buildings-parcels vector data adjustement by least squares. In Proc. 9th International Symposium on Spatial Data Quality (ISSDQ). **2015.**

Miller, G. A. WordNet: a lexical database for English. Communications of the ACM, **1995**, Vol. 38(11), pp. 39-41

Mochol, M. & Jentzsch, A. Towards a rule-based matcher selection. International Conference on Knowledge Engineering and Knowledge Management. Springer, Berlin, Heidelberg, **2008**

Mohri, M., Rostamizadeh, A. & Talwalkar, A. Foundations of machine learning. MIT press, **2012**

Moncla, L. Automatic reconstruction of itineraries from descriptive texts. Thèse de doctorat. Université de Pau et des Pays de l'Adour, Universidad de Zaragoza, **2015**

Monge, A. E. & Elkan, C. The Field Matching Problem: Algorithms and Applications. In KDD, **1996**, pp. 267-270

Mustière, S. & Devogele, T. Matching networks with different levels of detail. GeoInformatica, **2008**, Vol. 12(4), pp. 435-453.

Mustière, S. Apprentissage supervisé pour la généralisation cartographique (Doctoral dissertation), **2001**

Nentwig, M., Hartung, M., Ngonga Ngomo, A. C. & Rahm, E. A survey of current link discovery frameworks. Semantic Web, **2017**, Vol. 8(3), pp. 419-436.

Ngonga Ngomo, A. C. N. Orchid–reduction-ratio-optimal computation of geo-spatial distances for link discovery. In International Semantic Web Conference. Springer, Berlin, Heidelberg, **2013**(10), pp. 395-410

Ngonga Ngomo, A. C. N. & Auer, S. Limes-a time-efficient approach for large-scale link discovery on the web of data. In IJCAI, **2011**(7), pp. 2312-2317

Ngonga Ngomo, A.C. & Lyko, K. Eagle: Efficient active learning of link specifications using genetic programming. The Semantic Web: Research and Applications, **2012**, pp. 149-163

Ngonga Ngomo, A.C & Lyko, K. Unsupervised learning of link specifications: deterministic vs. non-deterministic. Proceedings of the 8th International Conference on Ontology Matching. CEUR-WS. org, **2013**, Vol. 1111

Ngonga Ngomo, A.C, Lyko, K. & Christen, V. Coala–correlation-aware active learning of link specifications. Extended Semantic Web Conference. Springer, Berlin, Heidelberg, **2013**

Nikolov, A., Uren, V., Motta, E. & De Roeck, A. Integration of semantically annotated data by the KnoFuss architecture. Knowledge Engineering: Practice and Patterns, **2008**, pp. 265-274

Niu, X., Rong, S., Zhang, Y. & Wang, H. Zhishi. links results for OAEI 2011. In Proceedings of the 6th International Conference on Ontology Matching. CEUR-WS. Org, **2011**, Vol. 814, pp. 220-227

Olteanu, A. Fusion de connaissances imparfaites pour l'appariement de données géographiques : proposition d'une approche s'appuyant sur la théorie des fonctions de croyance. Thèse de doctorat. Université Paris-Est, **2008**

Papadias, D. & Theodoridis, Y. Spatial Relations, Minimum Bounding Rectangles, and Spatial Data Structures. International Journal of Geographic Information Science, **1997**, Vol.11, pp. 111-138.

Pernelle, N., Saïs, F. & Symeonidou, D. An automatic key discovery approach for data linking. Web Semantics: Science, Services and Agents on the World Wide Web, **2013**, Vol. 23, pp. 16-30

Philips, L. Hanging on the metaphone. Computer Language, **1990**, Vol. 7(12)

Philips, L. The double metaphone search algorithm. C/C++ users journal, **2000**, Vol.18(6), pp. 38-43

Quodverte, P. La cartographie numérique et l'information géographique: importance et conséquences du progrès des sciences et des techniques. Thèse de doctorat. Orléans, **1994**

Raimond, Y., Sutton, C. & Sandler, M. B. Automatic Interlinking of Music Datasets on the Semantic Web. LDOW, **2008**, Vol. 369

Regnauld, N. Algorithms for the amalgamation of topographic data. Proceedings of the 21st International Cartographic Conference, Durban, South Africa, **2003**, Vol. 1016

Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007, **1995**

Ressler, J., Freese, E. & Boaten, V. "Semantic Method of Conflation". In: Terra Cognita 2009 Workshop In Conjunction with the 8th International Semantic Web Conference. Washington, USA, **2009**

Ruas, A. Modèle de généralisation de données géographiques à base de contraintes et d'autonomie (Doctoral dissertation), **1999**

Ruas, A. Généralisation et représentation multiple. Hermès science, **2002**

Russell, R. C. U.S. Patent No. 1,261,167. Washington, DC: U.S. Patent and Trademark Office, **1918**

Saïs, F., Pernelle, N. & Rousset, M. C. Combining a logical and a numerical method for data reconciliation. Journal on Data Semantics, **2009**, Vol. 12(12), pp. 66-94

Salas, J. & Harth, A. Finding spatial equivalences accross multiple RDF datasets. In Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web, **2011**, pp. 114-126

Samal, A., Seth, S. & Cueto, K. A feature-based approach to conflation of geospatial sources. IJGIS, **2004** (7-8), Vol. 18(5), pp. 459-489

Schade, S. Computer-tractable translation of geospatial data. International Journal of Spatial Data Infrastructures Research, Revue en ligne publiée par le Joint Research Centre (European Commission), **2010**, Vol. 5

Scharffe, F. & Euzenat, J. (2010, January). Méthodes et outils pour lier le Web de données. In Actes 17e conférence AFIA-AFRIF sur reconnaissance des formes et intelligence artificielle (RFIA), **2010**, pp. 678-685.

Scharffe, F., Liu, Y. & Zhou, C. Rdf-ai: an architecture for rdf datasets matching, fusion and interlink. In Proc. IJCAI 2009 workshop on Identity, reference, and knowledge representation (IR-KR), Pasadena (CA US), **2009**

Scharffe, F., Zhengjie, F., Ferrara, A., Khrouf, H. & Nikolov, A. Methods for automated dataset interlinking. Datalift Deliverable D4.1, HAL, **2011**

Seddiqui, M., Nath, R.P.D. & Aono, M. An efficient metric of automatic weight generation for properties in instance matching technique. arXiv preprint arXiv:1502.03556, **2015**

Sehgal, V., Getoor, L. & Viechnicki, P. D. Entity resolution in geospatial data integration. Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems. ACM, **2006**

Shafer, G. A. Mathematical Theory of Evidence. Princeton : Princeton University Press, **1976**

Sheeren, D. Méthodologie d'évaluation de la cohérence inter-représentations pour l'intégration de bases de données spatiales. Une approche combinant l'utilisation de métadonnées et l'apprentissage automatique. Thèse de doctorat. Université Pierre et Marie Curie-Paris VI, **2005**

Sherif, M. A., Dreßler, K., Smeros, P. & Ngomo, A. C. N. Radon-Rapid Discovery of Topological Relations. In AAAI, **2017**, pp. 175-181

Shvaiko, P. & Euzenat, J. Ontology matching: state of the art and future challenges. IEEE Transactions on knowledge and data engineering, Vol. 25(1), **,2013 ,** p.158-176.

Smeros, P. & Koubarakis, M. Discovering Spatial and Temporal Links among RDF Data. In LDOW@ WWW, **2016**

Smith, B. & Mark, D. Ontology and Geographic Kinds. 8th International Symposium on Spatial Data Handling (SDH'98), **1998**, pp. 308-320

Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. Journal of molecular biology, **1981**, Vol. 147(1), pp. 195-197

Soru, T., Marx, E. & Ngonga Ngomo, A.C. ROCKER: A refinement operator for key discovery. Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, **2015**

Sui, H., Li, D. & Gong, J. Automatic feature-level change detection (FLCD) for road network. In : Proceedings of the 20th ISPRS Congress, Istanbul, **2004**

Symeonidou, D., Armant, V., Pernelle, N. & Saïs, F. SAKey: Scalable almost key discovery in rdf data. In International Semantic Web Conference. Springer, Cham, **2014**(10), pp. 33-49

Taft, R. L. Name search techniques (No. 1). Bureau of Systems Development, New York State Identification and Intelligence System, **1970**

Touya, G. & Brando-Escobar, C. Detecting level-of-detail inconsistencies in volunteered geographic information data sets. Cartographica: The International Journal for Geographic Information and Geovisualization, **2013**, Vol.48(2), pp. 134-143

Troncy, R., Atemezing, G. A., Abadie, N. & Lam, C. V. Modeling geometry and reference systems on the web of data. In W3C Workshop on Linking Geospatial Data, **2014**

Uitermark, H. Ontology-based geographic data set integration. Thèse de doctorat. Université de Twente, **2001**

Vauglin, F. Modèles statistiques des imprécisions géométriques des objets géographiques linéaires. Thèse de Doctorat, Université Marne-la-Vallée, **1997**

Vidal, V., Sacramento, E., de Macêdo, J. & Casanova, M. An ontology-based framework for geographic data integration. Advances in Conceptual Modeling-Challenging Perspectives, **2009**, pp. 337-346

Vilches-Blázquez, L. M., Saquicela, V. & Corcho, O. Interlinking geospatial information in the web of data. Bridging the Geographic Information Sciences, **2012**, pp. 119-139

Vincenty, T. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Survey Review 22, **1975**, pp. 88-93

Volz, J., Bizer, C., Gaedke, M. & Kobilarov, G. Silk-A Link Discovery Framework for the Web of Data. LDOW, **2009**, Vol. 538

Volz, S. An iterative approach for matching multiple representations of street data. In ISPRS Workshop, Multiple representation and interoperability of spatial data, Hanover, Germany, **2006**, pp. 22–24

Walter, V. & Fritsch, D. Matching Spatial Data Sets: Statistical Approach. International Journal of Geographical Information Science, **1999**, Vol. 13(5), pp. 445-473

Wang, Y., Chen, D., Zhao, Z., Ren, F. & Du, Q. A Back-Propagation Neural Network-Based Approach for Multi-Represented Feature Matching in Update Propagation. Transactions in GIS, **2015**, Vol.19(6), pp. 964-993

Wienand, D. & Paulheim, H. Detecting incorrect numerical data in dbpedia. In European Semantic Web Conference. Springer, **2014**, pp. 504–518

Winkler, W. E. & Thibaudeau, Y. An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census. US Bureau of the Census, **1991**, pp. 1-22

Winkler, W. E. Overview of record linkage and current research directions. In Bureau of the Census, **2006**

World Wide Web Consortium. Ontologies. **2015**.
https://www.w3.org/standards/semanticweb/ontology

World Wide Web Consortium. RDF 1.1 concepts and abstract syntax. **2014.**
https://www.w3.org/TR/rdf11-concepts/

World Wide Web Consortium. W3c semantic web activity. **2008**. http://www.w3.org/2001/sw

Wu, Z. & Palmer, M. Verb Semantics and Lexical Selection, In: Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics, 1994, pp. 133-138

Yang, B., Luan, X. & Zhang, Y. A Pattern-Based Approach for Matching Nodes in Heterogeneous Urban Road Networks. Transactions in GIS, 2014, Vol.18(5), pp. 718-739.

Zaveri, A., Rula A., Maurino, A., Pietrobon, R., Lehmann, J. & Auer, S. Quality assessment for linked data: A survey. Semantic Web Journal, 2015

ANNEXES

Annexe A

Cette annexe représente Les requêtes Sparql effectuées pour l'insertion, en RDF, des métadonnées des géométries des adresses de la BD ADRESSE®. La liste suivante représente les préfixes des URis utilisées dans les requêtes. Les URIs de bases utilisées pour les données et le vocabulaire de la BD ADRESSE® sont des définies localement seulement pour notre cas d'application.

prefix geom:<http://data.ign.fr/def/geometrie#>.
prefix xys:<http://data.ign.fr/def/xysemantics#>.
prefix xysm:<http://data.ign.fr/id/codes/xysemantics/methodeevaluation/>.
prefix xyse:<http://data.ign.fr/id/codes/xysemantics/elementcaracteristique/>.
prefix bda:<http://localhost/source/bd_adresse-75/>.
prefix bdao:<http://localhost/def/bd_adresse-75#>.
prefix qudt:<http://qudt.org/schema/qudt#>.
prefix qudti:<http://www.qudt.org/qudt/owl/1.0.0/unit/Instances.html#>.
prefix xsd:<http://www.w3.org/2001/XMLSchema#>.
prefix dolce:<http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#>.

Insertion des métadonnées sur la précision planimétrique

Adresses modélisées aux plaques adresses

insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:absolutePositionalAccuaracy bda:accuracy1.

bda:accuracy1 a xys:AbsolutePositionalAccuracy.
bda:accuracy1 xys:evaluationResult bda:result1.

bda:result1 a xys:Result.
bda:result1 qudt:unit qudti#Meter.
bda:result1 qudt:value "12"^^xsd:float.
bda:accuracy1 xys:evaluationMethod bda:method1.

bda:method1 a xysm:SampleBasedInspection.
}
}
where {
?s a geom:Point.
?x geom:geometry ?s.
?x bdao:type_loc "Plaque adresse".
?x a bdao:Adresse.
}

Adresses modélisées par projection sur le tronçon

```
insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:absolutePositionalAccuaracy bda:accuracy2 .

bda:accuracy2 a xys:AbsolutePositionalAccuracy.
bda:accuracy2 xys:evaluationResult bda:result2.

bda:result2 a xys:Result.
bda:result2 qudt:unit qudti#Meter.
bda:result2 qudt:value "18"^^xsd:float.
bda:accuracy2 xys:evaluationMethod bda:method2.

bda:method2 a xysm:SampleBasedInspection.
}
}
where {
?x geom:geometry ?s.
?x bdao:type_loc "Projection".
?x a bdao:Adresse.
}
```

Adresses modélisées par interpolation sur la voie/ le tronçon

```
insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:absolutePositionalAccuaracy bda:accuracy3 .

bda:accuracy3 a xys:AbsolutePositionalAccuracy.
bda:accuracy3 xys:evaluationResult bda:result3.

bda:result3 a xys:Result.
bda:result3 qudt:unit qudti#Meter.
bda:result3 qudt:value "30"^^xsd:float.
bda:accuracy3 xys:evaluationMethod bda:method3.

bda:method3 a xysm:SampleBasedInspection.
}
}
where {
?x geom:geometry ?s.
?x bdao:type_loc "Voie".
?x a bdao:Adresse.
}
```

Adresses modélisées au centre de la commune / zone adressage

```
insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:absolutePositionalAccuaracy bda:accuracy4 .

bda:accuracy4 a xys:AbsolutePositionalAccuracy.
bda:accuracy4 xys:evaluationResult bda:result4.

bda:result4 a xys:Result.
bda:result4 qudt:unit qudti#Meter.
bda:result4 qudt:value "30"^^xsd:float.
bda:accuracy4 xys:evaluationMethod bda:method4.

bda:method4 a xysm:SampleBasedInspection.
}
}
where {
?x geom:geometry ?s.
?x a bdao:Adresse.
?x bdao:type_loc ?tl.
filter (?tl in ("Zone d'adressage","Commune"))
}
```

Insertion des métadonnées sur la modélisation géométrique

Adresses modélisées aux plaques adresses

```
insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:isModeledFrom bda:shapeelement1.
bda:shapeelement1 xys:shapeCharacteristicElementType xyse:FiatBoundaryPoint.
bda:shapeelement1 dolce:host topo:Bati.
}
}
where {
?x bdao:type_loc "Plaque adresse".
?s a geom:Point.
?x geom:geometry ?s.
?x a bdao:Adresse.
}
```

Adresses modélisées par interpolation sur le tronçon

```
insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:isModeledFrom bda:shapeelement2.
bda:shapeelement2 xys:shapeCharacteristicElementType xyse:CentrelinePoint.
bda:shapeelement2 dolce:host topo:Route.
}
where {
?x bdao:type_loc "Tronçon".
?s a geom:Point.
?x geom:geometry ?s.
?x a bdao:Adresse.
}
```

Adresses modélisées par interpolation sur la voie

```
insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:isModeledFrom bda:shapeelement2.
bda:shapeelement2 xys:shapeCharacteristicElementType xyse:CentrelinePoint.
bda:shapeelement2 dolce:host topo:Route.
}
where {
?x bdao:type_loc "Voie".
?s a geom:Point.
?x geom:geometry ?s.
?x a bdao:Adresse.
}
```

Adresses modélisées au centre de la zone adressage

```
insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:isModeledFrom bda:shapeelement3.
bda:shapeelement3 xys:shapeCharacteristicElementType xyse:Centroid.
bda:shapeelement3 dolce:host topo:ZoneAdressage.
}
where{
?x bdao:type_loc "Zone d'adressage".
?s a geom:Point.
?x geom:geometry ?s.
?x a bdao:Adresse.
}
```

Adresses modélisées au centre de la commune

insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:isModeledFrom bda:shapeelement5.
bda:shapeelement5 xys:shapeCharacteristicElementType xyse:Centroid.
bda:shapeelement5 dolce:host http://data.ign.fr/def/geofla#Commune*.*
}
where {
?x bdao:type_loc "Commune".
?s a geom:Point.
?x geom:geometry ?s.
?x a bdao:Adresse.
}

Adresses modélisées par projection sur le tronçon

insert {
graph<http://localhost/source/bd_adresse-75>{
?s xys:isModeledFrom bda:shapeelement4.
bda:shapeelement4 xys:shapeCharacteristicElementType xyse:ImplicitGeometricalBoundaryPoint.
bda:shapeelement4 dolce:host topo:Route.
}
where {
?x bdao:type_loc "Projection".
?s a geom:Point.
?x geom:geometry ?s.
?x a bdao:Adresse.
}