# Coalla : Un modèle pour l'édition collaborative d'un contenu géographique et la gestion de sa cohérence

Carmen Brando Escobar

UNIVERSITÉ PARIS-EST

**École Doctorale Mathématiques et**
**Sciences et Technologies de l'Information et de la Communication**

**THESE DE DOCTORAT**

pour obtenir le grade de

Docteur de l'Université Paris-Est

Spécialité : Informatique

Option : Sciences et Technologies de l'Information Géographique

présentée et soutenue publiquement par

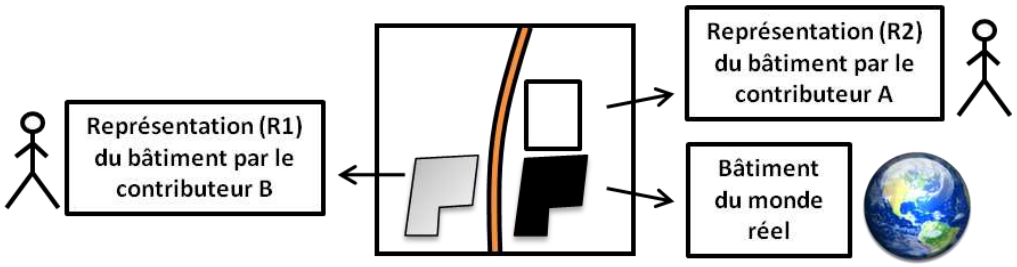
**Carmen BRANDO ESCOBAR**

le 5 avril 2013

encadrée et dirigée par

Bénédicte BUCHER

# *Coalla* : Un modèle pour l'édition collaborative d'un contenu géographique et la gestion de sa cohérence




**Composition du jury :**

| | |
|---|---|
| M. Thierry JOLIVEAU, Université de Saint-Etienne | Rapporteur |
| M. Jérôme GENSEL, Université de Grenoble II | Rapporteur |
| Mme. Anne DOUCET, Université de Paris VI | Examinateur |
| M. Ross PURVES, Université de Zurich | Examinateur |
| M. Gérôme CANALS, Université de Nancy II | Examinateur |
| Mme. Bénédicte BUCHER, COGIT-IGN | Encadrante et directrice de thèse |

*Cette thèse a été réalisée au Laboratoire COGIT de l'Institut National de l'Information Géographique et Forestière, sous la direction de Bénédicte Bucher.*

*Institut National de l'Information Géographique et Forestière*
*Service de la Recherche, Laboratoire COGIT*
*73 Avenue de Paris*
*94165 Saint-Mandé Cedex*
*Tél. : 01 43 98 80 00*

# Résumé

La production et la maintenance de contenus géographiques se fait souvent grâce à la mise en commun de contributions diverses. La mise à jour des données de l'IGN s'appuie ainsi sur l'intégration de données de partenaires ou la prise en compte d'alertes d'évolution du terrain. C'est également le cas des contenus libres produits par des projets communautaires comme Open Street Map.

Un aspect problématique est la gestion de la qualité d'un contenu géographique collaboratif, particulièrement de leur cohérence afin de permettre que des prises de décision s'appuient dessus. Cette cohérence est liée à l'homogénéité de la représentation de l'espace, ainsi qu'à la préservation d'informations importantes non explicites mais qui peuvent être retrouvées sur les entités décrites grâce à leurs géométries.

Ce travail de thèse propose un modèle baptisé *Coalla* pour l'édition collaborative d'un contenu géographique avec gestion de la cohérence. Ce modèle comporte trois contributions : 1) l'identification et la définition des éléments que doit comporter un vocabulaire formel visant à faciliter la construction d'un contenu géographique collaboratif ; 2) un processus d'aide à la construction à la volée d'un vocabulaire formel à partir de spécifications formelles des bases de données IGN et à des vocabulaires collaboratifs existants, et 3) une stratégie d'évaluation et de réconciliation des contributions afin de les intégrer d'une façon cohérente au contenu central. Notre modèle Coalla a été implémenté dans un prototype.

**Mots-clés :** processus collaboratifs de production, données géographiques communautaires, cohérence de données géographiques, relations spatiales, vocabulaires formels, réconciliation des contributions d'utilisateurs

# Abstract


***Coalla*: A Model for Collaborative Editing and Consistency Management of Geographic Content**

Geographic content production and maintenance is often done through a combination of various contributions. Thus, updating IGN geographic data relies on integrating data from partners or by involving field change alerts. This is also the case of free content produced by community projects such as OpenStreetMap

An important problem is quality management of collaboratively produced geographic content, in particular consistency management. This allows for decision-making which is based on this content. Data consistency depends on how homogenous space representation is. Likewise, it depends on preserving important non-explicit information that can be found on the geometries of the entities described in the content.

This Thesis proposes a model baptized *Coalla* for collaborative editing of geographic content with consistency management. The model has three contributions: 1) identifying and defining elements that should be included in a formal vocabulary to facilitate the construction of collaborative geographic content, 2) user assistance process to help users build on the fly a formal vocabulary extracted from formal IGN databases specifications and existing collaborative vocabularies, and 3) a strategy for evaluating and reconciling user contributions in order to coherently integrate them into the content. Our model *Coalla* has been implemented in a prototype.



**Keywords:** collaborative production processes, volunteered geographic information, geographic data consistency, spatial relations, formal vocabularies, reconciliation of user contributions

# Remerciements

Je remercie tout d'abord à ma directrice et encadrante de thèse Bénédicte Bucher. Son soutien scientifique et moral tout au long de ces trois années a été indispensable pour réussir mon objectif. Bénédicte, merci encore pour tes encouragements et ton optimisme.

Merci aux membres du jury, M. Thierry Joliveau, M. Jérôme Gensel, Mme. Anne Doucet, M. Ross Purves et M. Gêrome Canals qui m'ont fait l'honneur d'éxaminer ce mémoire. L'échange que nous avons eu lors de la phase de questions de la soutenance était très enrichissant.

Je tiens à remercier à Anne Ruas et Sébastien Mustière de m'avoir très bien accuellie au COGIT de même que pour leurs remarques et leurs conseils pendant les réunions du laboratoire et mes répetitions de présentations. Marie Claude Foubert du Service de la Recherche de l'IGN et Sylvie Cach de l'Université Paris-Est, vous m'avez beaucoup aidé pendant ma vie de doctorante. Je remercie à Guillaume Touya pour sa bonne energie, son intérêt pour mon travail et son expertise inmense. Merci à mes relecteurs fidèles dans l'ordre de relecture du mémoire : Nicolas Lebas, Eric Mermet, Guillaume Touya et Nathalie Abadie. Merci aussi à mon ancienne stagiaire Nassima Chenachena pour son travail et les bons échanges. Merci à mes collèges du COGIT pour les discussions, les bons moments, les encouragements, Jef, Catherine, Kusay, Elodie, Sidonie, Firas, Jérémy, Fayçal, Julien, …

Un gran merci aux contributeurs d'OpenStreetMap, en particulier Christian Quest, pour leur participation au sondage, pour les cartoparties et les retours. Merci aux opérateurs de la MAJEC de l'IGN de me permettre me soumerger dans les activités de collecte et de saisie de données. Merci aussi à Bruno Bordin et Frank Fuchs du service de développement de l'IGN pour les échanges.

Merci à mes bonnes amies, Léa Massiot avec qui j'ai beaucoup discuté sur Wikipédia, merci pour les beaux moments d'amitié. Leidiana Martinis, ton esprit gai m'a aidé pendant la rédaction. Laurence Jolivet, je pouvais toujours te parler, merci pour m'écouter. Christine Plumejeaud, merci pour tes encouragements, les moments très sympa à midi et après la thèse sur Paris. Mes amis et collèges de bureau, les deux Erics (Eric Mermet et Eric Grosso), sans vous deux l'expérience de la thèse n'aurait pas été si spéciale.

Je tiens aussi à remercier aux vespuccians de l'école d'été Vespucci (édition 2011), en particulier aux conférenciers et organisateurs Michael Goodchild, Muki Haklay et bien spécialement à Werner Kuhn de l'Université de Münster, aussi pour les discussions à AGILE et GISRUK. Ces expériences ont été très enrichissantes et m'ont permises d'explorer différents points de vue dans mes recherches.

Merci aux laboratoires CRHM et PIREH de l'Université Paris 1, en particulier Mme. Christine Lebeau, Stéphane Lamassé et Julien Alerini pour leurs encouragements pendant la préparation de la soutenance.

Gracias a mis padres, Vincenzo y Maria, Uds. son ejemplo de trabajo et perserverancia. Gracias por apoyarme en mis metas y de darme animos los domingos en la noche por telefono. Gracias a mi hermana Angela et mi hermano Vicente por sus buenos deseos y ayuda incondicional.

Nicolas, grâce à toi, j'ai pu tenir pendant les mois de la rédaction.

*A mis padres.*

# Table des matières

# Introduction

## 1 Contexte et objectifs

La production et la maintenance de contenus géographiques se fait souvent grâce à la mise en commun de contributions diverses. Ces activités collaboratives posent fréquemment certains problèmes concernant les différents points de vue, d'expertise et de capacité d'observation des individus qui participent à l'activité d'édition des données. Ces problèmes impactent la qualité des données géographiques produites et plus particulièrement leur cohérence.

En information géographique, la cohérence d'un contenu est liée d'une part à l'homogénéité de la représentation de l'espace. Sachant qu'une représentation ne clone pas l'espace géographique mais ne peut qu'en restituer des caractéristiques vues sous un certain filtre, il est important que ce filtre soit homogène sur le territoire couvert pour ne pas induire l'utilisateur à faire de fausses interprétations. La cohérence d'un contenu géographique est également liée à la préservation d'informations importantes non explicites. Cependant, ces informations peuvent être retrouvées sur les entités décrites grâce à leurs géométries. Cette cohérence représente souvent le respect de règles de sens commun qui implique ne pas avoir des incohérences dans la base de données. Par exemple, une maison qui se superpose à une route est une incohérence topologique.

Aujourd'hui, il existe un besoin de faciliter l'édition collaborative d'un contenu géographique par des individus et de les aider à gérer la cohérence de ce contenu afin de permettre que des prises de décision s'appuient dessus. Ce besoin est réel dans deux contextes de production de données géographiques de natures différentes. D'une part, la mise à jour des données de l'IGN s'appuie ainsi sur l'intégration de données de partenaires ou la prise en compte d'alertes d'évolution du terrain. D'autre part, les projets communautaires de cartographie collaborative comme Open Street Map (OSM) permettent aux contributeurs de constituer des contenus géographiques, comme par exemple, les arrêts et les lignes de bus sur une ville. Le projet OSM est un des meilleurs exemples du phénomène récent baptisé *Volunteered Geographic Information* (VGI) où les citoyens deviennent capteurs des changements de leurs espaces grâce aux nouvelles technologies.

Dans ces contextes, l'objectif de ce travail de thèse est d'aider les individus à constituer et gérer la cohérence de leur contenu géographique collaboratif.

# 2 Proposition

Pour atteindre nos objectifs, nos recherches s'appuient sur plusieurs idées principales.

La première idée est que, dans un contexte de production communautaire, les utilisateurs puissent raccrocher le contenu collaboratif à un contenu de référence afin d'améliorer la cohérence et donc l'utilisabilité du contenu. Un des intérêts de l'utilisateur est de pouvoir signaler des caractéristiques de certains objets du contenu, par exemple, « c'est le plus grand immeuble de la zone ». Il peut être aussi intéressé d'indiquer des relations typées entre les objets du contenu et les objets du contenu de référence, par exemple, « l'abribus du contenu collaboratif est exactement en face du bureau de poste de référence ». De plus, il peut aussi vouloir définir des contraintes d'intégrité qu'il veut voir respectées dans son contenu, à partir de types de relations spatiales (ex : chevaucher) entre les classes du contenu et les concepts du contenu de référence, par exemple, « les bâtiments du contenu collaboratif ne chevauchent généralement pas les routes RGE® ».

La deuxième idée est qu'en l'absence de spécifications ou de règles dans un contexte de production communautaire, il faut aider le contributeur à construire un vocabulaire formel afin de décrire les entités, leurs propriétés spatiales et leurs relations spatiales avec d'autres entités, et ainsi assurer une certaine homogénéité de la représentation de l'espace.

La troisième idée est que les contributions peuvent être intégrées de façon à obtenir un contenu cohérent de deux manières qui se complémentent. La première façon est par l'évaluation de contraintes d'intégrité afin de détecter les incohérences liées à la non-préservation des relations importantes entre le contenu collaboratif et le contenu de référence. La deuxième manière est la réconciliation d'éditions provenant de contributeurs différents. Plus précisément, cette réconciliation prend en compte la décomposition du contenu en considérant la notion de structure afin de rendre indépendantes les éditions sur des parties différentes du contenu.

En considérant ces idées, nous proposons un modèle d'édition collaborative d'un contenu géographique pour la gestion de sa cohérence qui s'appuie sur un vocabulaire formel et des contraintes d'intégrités, de même que sur un modèle d'édition et des stratégies d'intégration des éditions afin de préserver la cohérence des données.

# 3 Plan du mémoire

Ce mémoire est constitué de quatre chapitres.

Le chapitre I décrit les deux contextes de production collaborative sur lesquels cette thèse est inscrite. Nous avons investigué les ressemblances et les différences entre ces deux contextes de production de données géographiques de nature différente : l'IGN et le projet OSM.

Le chapitre II correspond à l'état de l'art. Ici, nous nous posons la question de ce « que veut dire la cohérence » dans différents types de contenus (géographiques et collaboratifs). Pour cette raison, nous investiguons des travaux existants qui se sont intéressés à la gestion de la qualité et la cohérence de trois types de contenu : des contenus non-géographiques dans des éditeurs collaboratifs, des données géographiques communautaires dites VGI, des données géographiques « traditionnelles ».

Le chapitre III présente notre modèle pour l'édition collaborative d'un contenu géographique et la gestion de sa cohérence baptisé ***Coalla***, pour souligner le mot « collaboration ». Ce chapitre décrit également une méthode d'aide à la construction d'un vocabulaire formel et nos stratégies d'intégration des contributions dans un contenu géographique commun.

Le chapitre IV correspond à la mise en œuvre de nos contributions.

Enfin, nous présentons nos conclusions et les perspectives de ce travail.

# Chapitre I

# Contexte : la production collaborative de données géographiques et la gestion de leur qualité

Ce chapitre décrit les processus de production collaborative de données géographiques et de gestion de leur qualité depuis les points de vue de deux mondes de natures différentes. D'un côté, les producteurs nationaux de données géographiques comme l'IGN sont chargés d'assurer la production, l'entretien et la diffusion de l'information géographique de référence. De l'autre côté, l'essor des technologies du Web 2.0 a permis la constitution de projets communautaires en ligne visant à créer et à mettre gratuitement à disposition des données géographiques libres. Aujourd'hui, le projet communautaire le plus populaire est le projet de cartographie collaborative Open Street Map (OSM) qui offre depuis 2004 une carte du monde éditable en ligne.

La section 1 explique les notions de base sur la qualité des données géographiques. La description de ces notions est essentielle pour comprendre les mécanismes de gestion de la qualité des données géographiques du point de vue d'un producteur institutionnel de données et d'une communauté de cartographie collaborative.

La section 2 décrit la production collaborative de données géographiques du point de vue d'un producteur traditionnel comme l'IGN et la manière dont la qualité d'un tel contenu est gérée. Le référentiel de données géographiques RGE® a été constitué entre 2000 et 2008 et depuis, les opérateurs de l'équipe MAJEC, acronyme de « Mise à jour en continu », sont chargés des activités de mise à jour des bases de données IGN. La production collaborative s'appuie sur les pratiques suivies par ces opérateurs, ainsi que les spécifications de contenu de ces bases de données, car elles permettent la construction d'un produit homogène et cohérent bien que la production soit le fait de plusieurs opérateurs. La production collaborative à l'IGN s'appuie aussi sur une architecture client/serveur et leurs échanges, plus précisément sur les clients d'édition des données, la structure de la base de données centrale BDUni, acronyme pour « Base de données unifiée », la façon dont les évolutions sont stockées sous forme d'historique dans cette base, et les mécanismes du côté serveur qui aident à gérer la qualité des données. Enfin, un

des éléments clé de la gestion de la qualité est la procédure d'évaluation effectuée par l'unité de qualité de l'IGN qui suit les spécifications ainsi que la documentation existante sur les évaluations de qualité.

La section 3 décrit la production collaborative d'un contenu géographique sur le Web 2.0, du point de vue d'un projet communautaire de cartographie collaborative comme OSM, et la façon dont la qualité d'un tel contenu est gérée. La production collaborative d'OSM s'appuie sur les consignes dictées par la communauté dans le site de documentation en ligne Wiki OSM que les contributeurs sont encouragés à respecter. Cette production repose également sur les pratiques de contribution, les clients d'édition, la structure de la base de données centrale OSM, la gestion de son historique et les messages d'échanges entre les clients et le serveur via l'API OSM, ces processus aidant globalement à gérer la qualité des données. Étant donné qu'il n'y a pas de restrictions rigides sur ce qui peut être saisi dans OSM, la communauté a mis en place des outils externes pour la détection des incohérences dans les données. Cela permet donc de détecter, dans une certaine mesure, des erreurs introduites lors de l'édition faite par plusieurs contributeurs et par l'absence de spécifications.

La section 4 présente une discussion comparant les points de ressemblance et de divergence de la production collaborative de données géographiques du point de vue d'un producteur institutionnel comme l'IGN et du point de vue d'un projet communautaire de cartographie en ligne comme OSM.

# 1 Quelques notions de base sur la qualité des données géographiques

Dans cette section, nous rappelons des notions de bases sur la qualité des données géographiques. C'est une discipline centrale et une des plus anciennes dans les sciences de l'information géographique. La qualité des données géographiques est traditionnellement observée selon deux points de vue, celui du producteur des données, appelée qualité interne, et celui de l'utilisateur de ces données, appelée qualité externe (Vauglin 1997).

Concernant la qualité interne, le producteur décide de représenter un sous-ensemble du monde réel dans les données. Comme le rappelle Bucher (2011), le contenu ne peut pas viser à reproduire à l'identique l'espace mais plutôt, il doit être fidèle à des choix d'observation qui ont été faits sur l'ensemble d'un territoire, c'est-à-dire, être fidèle au terrain nominal. Ces choix effectués par le producteur sont explicités dans les spécifications ; elles décrivent ce que les données auraient dû être si elles étaient parfaites (Devillers et Jeansoulin 2006). Par exemple, un producteur peut choisir d'utiliser un filtre comme « les pistes cyclables ayant des longueurs inférieures à 200 mètres sont exclues ». Ces pistes cyclables ne seront donc pas représentées dans la base de données. La qualité interne est évaluée à l'aide de plusieurs indicateurs qui mesurent l'écart entre les données produites et le terrain nominal. Ces indicateurs sont des métadonnées standards de qualité définis dans la norme ISO 19115 (ISO 2003) dont les éléments sont les suivants : la généalogie, la précision géométrique, la précision thématique, la précision temporelle, la cohérence logique, et l'exhaustivité. La précision géométrique pour certains objets topographiques dépend du thème auquel ils appartiennent et de la source ayant

servi à la saisie. Par exemple, la précision planimétrique (2D) d'un objet appartenant à un des thèmes « réseau routier, voies ferrées, hydrographie terrestre, réseau de distribution, bâtiments, végétation » est affectée avec une valeur de 2,5 mètres pour les levés GNSS.

La qualité externe définit le niveau d'adéquation entre les spécifications et les besoins des utilisateurs des données (Devillers et Jeansoulin 2006). Un utilisateur des données devrait être capable d'évaluer leur utilisation en regardant un extrait des données, les métadonnées et les spécifications afin de savoir quels aspects du monde réel ont été représentés dans les données, et les écarts avec le terrain nominal. Fréquemment, le producteur communique les possibles usages du contenu dans les spécifications en présentant plusieurs cas d'applications. Ces ressources mises à disposition pour l'utilisateur peuvent être très complexes (même pour un expert). En conséquence, il est difficile pour lui de comprendre l'information présentée et quelle est la pertinence pour son application (Devillers et al. 2004; Fisher et al. 2009). Pour cette raison, diverses approches visent à formaliser les besoins de l'utilisateur dans une ontologie afin d'évaluer automatiquement la qualité externe (Vasseur et al. 2005). Les auteurs comparent cette ontologie avec une autre ontologie décrivant les données, et la qualité externe correspond donc au degré de similarité entre les caractéristiques des données et les besoins de l'utilisateur. Par exemple, dans le contexte d'une application d'aide à la navigation en transport urbain pour les touristes, il est important d'indiquer que pour chacun des « points d'intérêt », le groupe de touristes souhaitera obtenir les qualités avec les tolérances suivantes : actualité (≥2002), exhaustivité (≥95 %), et exactitude de positionnement (≤1 mètre). La qualité externe est aussi concernée par les techniques de géovisualisation qui sont des moyens très intéressants pour communiquer de la qualité avec d'autres utilisateurs (Bucher 2009). (Mackaness et al. 1994) compilent une large sélection de travaux dans ce domaine. Plus récemment, le mode de production communautaire favorise la qualité externe dans la mesure où une communauté peut décider de construire ensemble précisément la base de données dont elle a besoin, comme les agressions relevées dans un quartier dans SpotCrime[1] (Bucher 2011). Un autre aspect associé à la qualité externe, est la confiance de l'utilisateur dans le contenu. Cet aspect semble être directement lié à la crédibilité du producteur, semblable à l'indicateur « légitimité » défini dans la norme ISO 19113 (ISO 2002) qui permet d'évaluer la crédibilité d'une donnée. Cette dimension de la qualité devient très difficile à gérer dans le contexte de la production d'un contenu communautaire (Bishr et Kuhn 2007) : une donnée produite par un producteur traditionnel de données paraît plus fiable que celle produite par des contributeurs dont les origines et compétences sont inconnues.

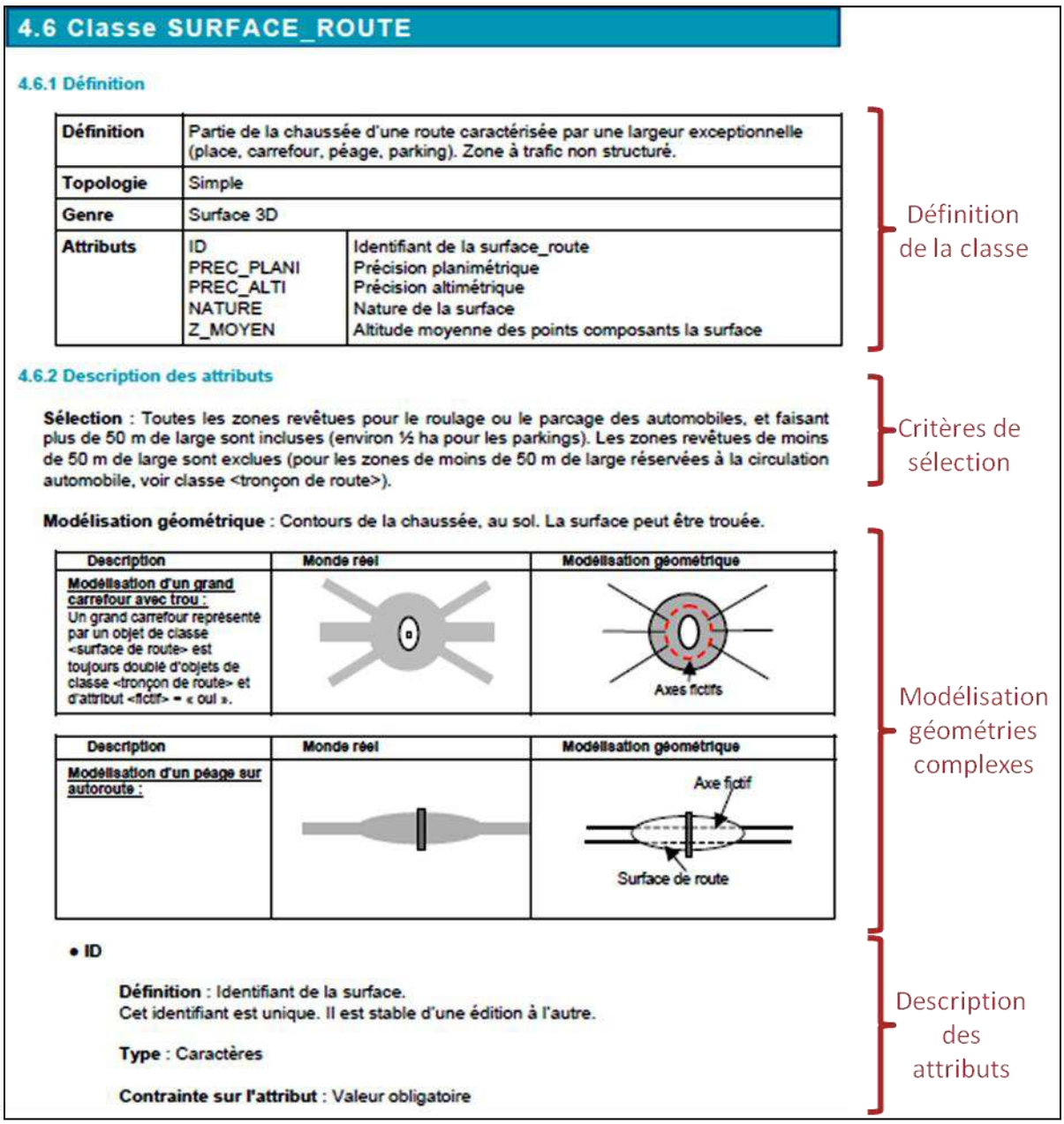
# 2 La production traditionnelle des données géographiques à l'IGN et la gestion de leur qualité

Les opérateurs de l'équipe MAJEC sont chargés des activités de mise à jour du RGE® (constitution achevée fin 2008). Cette section décrit donc ce processus qui s'appuie sur des spécifications.

[1] http://www.spotcrime.com/

## 2.1 Les spécifications

Les spécifications des bases de données IGN permettent d'assurer une observation homogène du monde réel bien que la production soit le fait de plusieurs opérateurs. Ces spécifications sont globalement organisées d'une façon similaire à ce que propose la norme ISO 19109 (ISO 2005). La Figure 1 illustre un extrait des spécifications de la BDTopo® de l'IGN correspondant à la classe Surface Route.



**4.6 Classe SURFACE_ROUTE**

**4.6.1 Définition**

| | | |
|---|---|---|
| **Définition** | Partie de la chaussée d'une route caractérisée par une largeur exceptionnelle (place, carrefour, péage, parking). Zone à trafic non structuré. | |
| **Topologie** | Simple | |
| **Genre** | Surface 3D | |
| **Attributs** | ID<br>PREC_PLANI<br>PREC_ALTI<br>NATURE<br>Z_MOYEN | Identifiant de la surface_route<br>Précision planimétrique<br>Précision altimétrique<br>Nature de la surface<br>Altitude moyenne des points composants la surface |

Définition de la classe

**4.6.2 Description des attributs**

**Sélection** : Toutes les zones revêtues pour le roulage ou le parcage des automobiles, et faisant plus de 50 m de large sont incluses (environ ½ ha pour les parkings). Les zones revêtues de moins de 50 m de large sont exclues (pour les zones de moins de 50 m de large réservées à la circulation automobile, voir classe <tronçon de route>).

Critères de sélection

**Modélisation géométrique** : Contours de la chaussée, au sol. La surface peut être trouée.

| Description | Monde réel | Modélisation géométrique |
|---|---|---|
| **Modélisation d'un grand carrefour avec trou :** Un grand carrefour représenté par un objet de classe <surface de route> est toujours doublé d'objets de classe <tronçon de route> et d'attribut <fictif> = « oui ». | | Axes fictifs |

| Description | Monde réel | Modélisation géométrique |
|---|---|---|
| **Modélisation d'un péage sur autoroute :** | | Axe fictif<br>Surface de route |

Modélisation géométries complexes

- **ID**

**Définition** : Identifiant de la surface.
Cet identifiant est unique. Il est stable d'une édition à l'autre.

**Type** : Caractères

**Contrainte sur l'attribut** : Valeur obligatoire

Description des attributs

**Figure 1 : Spécification de la classe Surface Route de la BDTopo® (IGN 2011c)**

Cette spécification décrit en détail le modèle de données en déclarant une fiche par classe qui comprend, selon (Gesbert 2005): « *le nom de la classe, sa définition générale, les éventuelles*

*relations la concernant, les critères de sélection restreignant la définition générale pour décrire précisément l'ensemble des entités que la classe est censée représenter, puis les attributs et leurs descriptions »*. Plus précisément, la spécification décrit des critères de sélection, par exemple « *les pylônes et portiques soutenant des lignes de 63 KV sont inclus* » (IGN 2011c) (p. 56). La spécification détaille aussi la modélisation des géométries pour les objets complexes afin qu'ils puissent être saisis de manière cohérente, par exemple, la règle pour saisir certains bâtiments : « *plusieurs bâtiments contigus de la même nature sont considérés comme un seul et même objet, seul le contour extérieur est saisi* » (IGN 2011c) (p. 78). La spécification indique aussi certaines relations importantes entre des objets de différents types, par exemple une règle explicitant une relation entre des objets de la classe Transport par Câble et des objets de la classe Pylône est indiquée ainsi dans les spécifications de la classe Transport par Câble : « *ligne brisée joignant le sommet de chaque pylône constituant un support de la ligne* » (IGN 2011c) (p. 45).

## 2.2 Le processus de mise à jour en continu

Les activités de mise à jour de la base de données centrale BDUni (dont les composantes adresse et topographique du RGE® sont dérivées), sont effectuées par l'équipe MAJEC. Elle est composée de six unités, chacune correspondant à une grande zone de la France métropolitaine (voir la Figure 2) : Nord-Est, Ile-de-France, Nord-Ouest, Centre-Est, Sud-Ouest, et Sud-Est. Chacune de ces unités possède un responsable et entre 1 et 5 opérateurs par département. Il faut noter qu'un même opérateur peut appartenir à plusieurs unités et donc s'occuper de plusieurs départements. Pour mieux comprendre les pratiques des opérateurs de la MAJEC, nous avons participé à une activité de collecte et de sasie de données, le 8 et 13 septembre 2010, respectivement. La tâche principale d'un opérateur est d'acquérir les évolutions sur sa zone, appelée zone de travail. Une évolution est un changement observable et observé sur le paysage (Badard 2000). Il relève également de l'opérateur de détecter les erreurs observées par rapport à l'état actuel de la BDUni qui sont souvent des oublis d'un objet lors de la saisie, un mauvais renseignement d'une valeur d'attribut ou une mauvaise localisation d'un objet (Badard 2000). Enfin, il incorpore ces évolutions en éditant le contenu sous la forme des créations, suppressions, modifications d'objets sur le contenu de la base de données concernée.

Autour de la restitution photogrammétrique, il existe principalement deux méthodes d'acquisition des évolutions et des erreurs suivies par un opérateur. Premièrement, les documents numériques mis à disposition, parfois en ligne sur l'Internet, par les partenaires de l'IGN comme les mairies, les pompiers et les conseils généraux, sont exploités manuellement par les opérateurs afin d'obtenir en quasi totalité les attributs thématiques sur certains objets comme les zones d'intérêt ou d'activité. Deuxièmement, le levé topographique permet d'acquérir les nouvelles géométries des objets sur le terrain à l'aide d'un récepteur d'antenne GNSS[2], ce qui est utile pour les objets les plus évolutifs comme les giratoires ou les bretelles d'autoroutes qui sont particulièrement difficiles à acquérir par la restitution photogrammétrique. Ce mode est également utilisé pour l'acquisition d'objets comme certains chemins de montagne dans des zones très arborées difficiles à obtenir par restitution. L'opérateur peut se déplacer en deux modes, soit en mode piéton (à pied ou à vélo) ou soit en mode embarqué (en voiture). L'opérateur doit auparavant préparer sa sortie sur le terrain. Il doit optimiser ses déplacements,

[2] Acronyme pour *Global Navigation Satellite* System, un récepteur d'antenne GNSS reçoit les signaux des satellites du système de localisation global. En présence d'un environnement obstrué, ces signaux sont souvent amplifiés grâce à une antenne incorporée.

et pour cette raison, il organise ses itinéraires en choisissant les endroits qui ont le plus probablement changé. L'opérateur garde toujours contact par téléphone ou par mail avec une personne appelée « correspondant ». Cette personne appartient à une préfecture, un conseil général, ou un autre organisme national ou régional qui est directement concernée par l'évolution d'un objet. L'opérateur surveille aussi les zones dites évolutives qui indiquent par exemple, la présence de chantiers sur le terrain. Chaque zone évolutive est représentée comme un objet surfacique englobant l'emplacement du chantier. Ces objets spéciaux ont des attributs particuliers comme la date (estimée) de finalisation du chantier, saisie par l'opérateur que l'on retrouve généralement dans les documents publics décrivant le projet de construction. Le logiciel client déclenche automatiquement des alertes à l'opérateur lorsque la période de temps indiquée est passée. Afin de préparer la sortie sur le terrain, l'opérateur doit également vérifier la présence suffisante de satellites sur la zone géographique à changer. Une fois sur le terrain, l'opérateur se sert d'un client léger basé sur le logiciel propriétaire GeoConcept, installé sur une Tablette PC contenant une copie locale de la dernière version de la BDUni, correspondant à la zone à changer et les types d'objets concernés. Il effectue l'enregistrement de la trace GNSS permettant d'acquérir les positions brutes au format .SSF. Suite à l'acquisition, l'opérateur effectue au bureau un post-traitement de ces traces afin d'améliorer leur précision de l'ordre du centimètre. L'opérateur se sert du réseau GNSS permanent (RGP), mis en place par l'IGN, composé de stations de référence fournissant des corrections en temps réel.

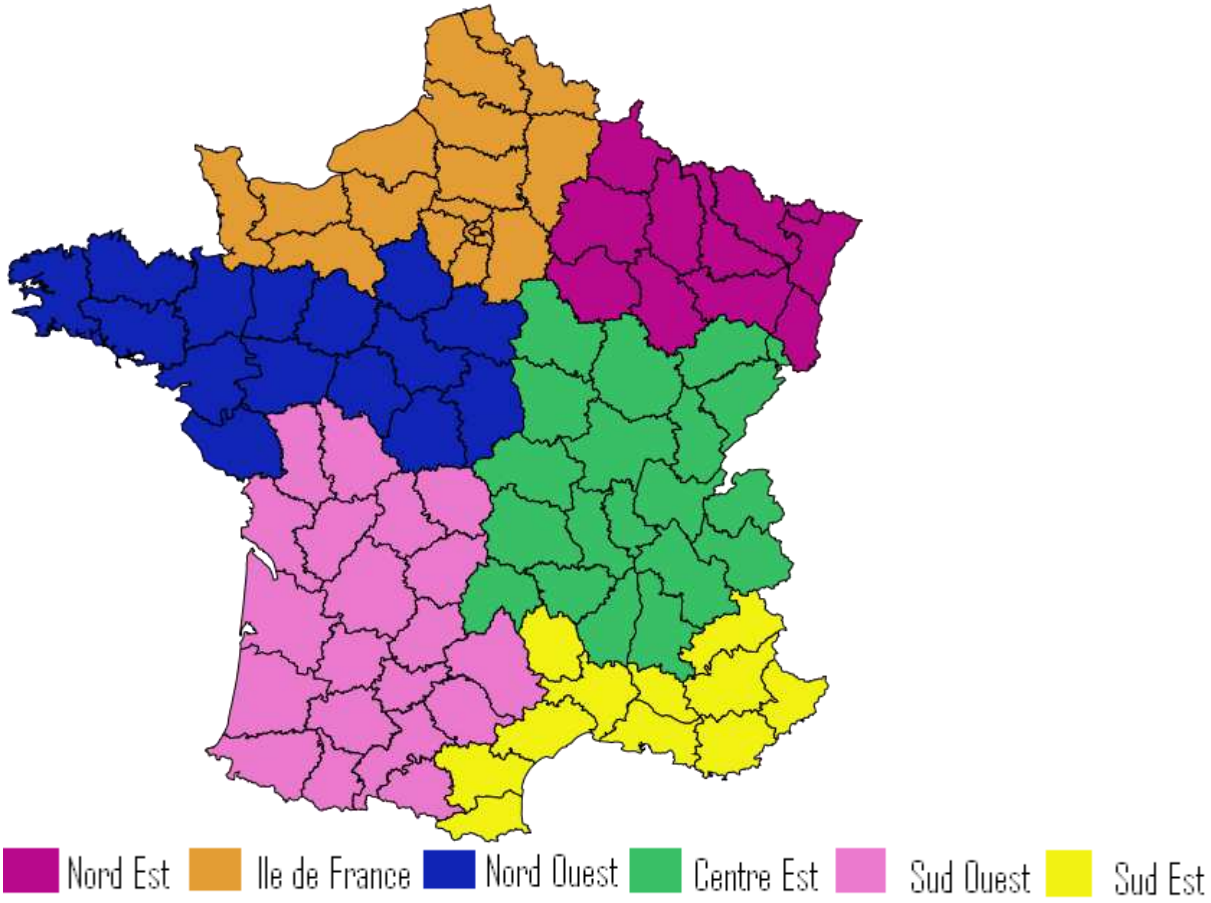



**Figure 2 : Distribution des opérateurs de la MAJEC en France Métropolitaine, les chiffres en noir, blanc, et jaune indiquent respectivement, les départements partagés entre plusieurs collecteurs, d'un seul opérateur, et d'aucun opérateur (IGN 2006)**

Ensuite, l'opérateur effectue l'édition du contenu de la BDUni correspondant à sa zone de travail afin d'incorporer les évolutions précédemment acquises. Le travail d'édition s'appuie sur un système client/serveur. Un serveur qui gère la BDUni, son historique, et les postes clients hors-ligne sur lesquels les opérateurs effectuent la saisie. Un poste client est composé du logiciel propriétaire SIG GeoConcept qui sert d'appui visuel à l'opérateur, du logiciel propriétaire GCVS, chargé d'établir la communication avec le serveur, et d'une copie locale de la zone de travail indépendante de la BDUni (la copie locale n'est pas synchronisée en permanence avec le contenu de la BDUni). Le processus d'échange d'une version d'un contenu de la BDUni correspondant à une zone spécifique d'un client vers le serveur est connu sous le nom de réconciliation[3]. Préalablement à ce processus, l'opérateur dessine une surface sur la carte sous GeoConcept afin de définir une zone dite de réconciliation qui sert à limiter la quantité d'objets à rechercher sur la BDUni par le serveur. Cette zone doit donc être assez large pour englober toutes les modifications mais pas très large non plus pour que la réconciliation ne prenne pas trop de temps (IGN 2006). L'opérateur peut incorporer les évolutions sur la zone de réconciliation. Ensuite, l'opérateur déclenche dans le client, les procédures de contrôle qui effectuent des opérations sur la copie locale du contenu afin de détecter par exemple des incohérences topologiques ou des doublons. Chacune de ces méthodes est implémentée dans GeoConcept (voir la Figure 3) par une équipe de développement à l'IGN, et sont fondamentales pour gérer la qualité de la BDUni.

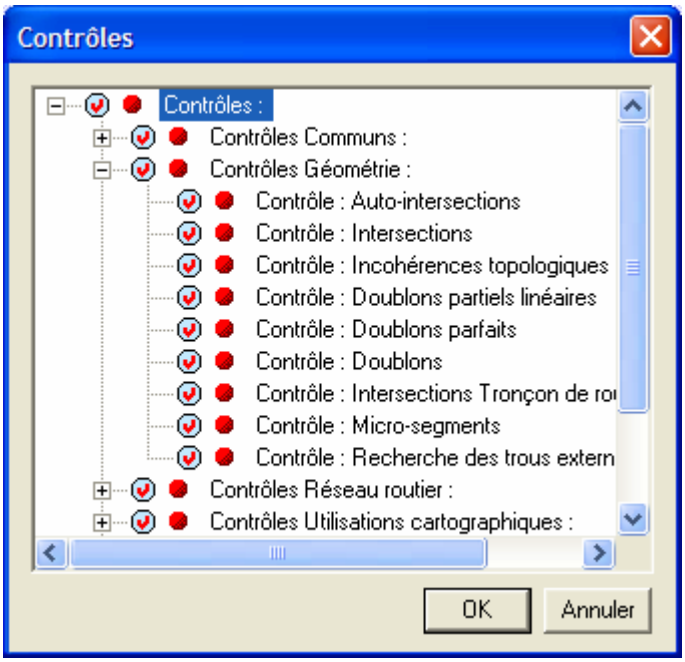



**Figure 3 : les procédures de contrôle à déclencher par l'opérateur afin de vérifier les incohérences de sa copie locale du contenu (IGN 2006)**

Une fois les incohérences corrigées manuellement, l'opérateur déclenche le processus de réconciliation décrit dans le diagramme de séquence UML de la Figure 4. La réconciliation doit assurer que tout objet modifié sur le serveur parviendra au client et que tout objet modifié sur le client parviendra au serveur (IGN 2006). Un processus de réconciliation est constitué de deux

[3] Le terme « réconciliation » (aussi utilisée dans la sous-section suivante), est utilisé plus tard dans le manuscrit pour nommer un concept différent du sens décrit ici (voir section III.3.2).

phases. Premièrement, le serveur transfère vers le client les objets étant dans la zone de réconciliation (et autour) qui ont été modifiés auparavant dans la BDUni par d'autres opérateurs. L'opérateur peut toujours voir les modifications effectuées par les collecteurs voisins concernant sa zone. Deuxièmement, le client transfère vers le serveur, les objets modifiés par lui même se trouvant dans la zone de réconciliation. Le serveur recherche les conflits de versions liées à la suppression et la modification : les objets détruits sur le serveur et modifiés sur le client, les objets modifiés sur le serveur et modifiés sur le client, et les objets modifiés sur le serveur et détruits sur le client. En absence de conflits, le serveur transforme les modifications sous la forme de requêtes DML[4] SQL et les exécute dans une transaction de base de données afin d'assurer l'intégrité de la BDUni. En présence de conflits, l'opérateur est notifié des conflits, et les objets concernés vont alors exister sous deux formes dans le client : la forme du serveur et la forme du client avec un marqueur signalant le conflit (IGN 2006). La résolution d'un conflit de versions a toujours besoin de l'intervention humaine. En cas d'édition d'un même objet, les collecteurs se mettent d'accord entre eux pour résoudre tel conflit, par exemple choisir une valeur pour un attribut entre deux valeurs différentes. L'opérateur est ensuite chargé de réappliquer ses changements sur la version serveur des objets. Ces nouvelles versions des objets et les informations sur la réconciliation sont stockées dans la BDUni, de même que les anciennes versions de tels objets sous la forme d'historique.

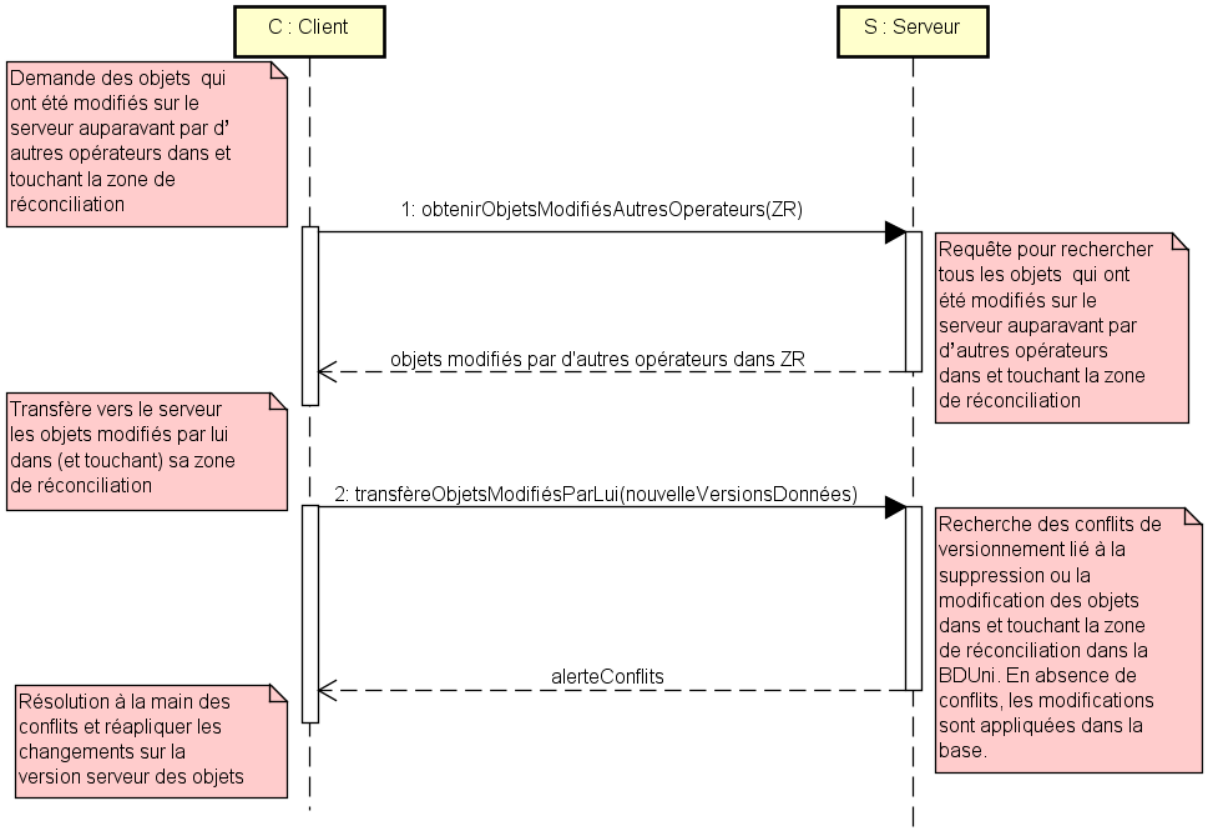



**Figure 4 : Diagramme de séquence UML pour décrire le processus de réconciliation à l'IGN**

[4] En anglais *Data Manipulation Language* pour les requêtes SQL de type INSERT, UPDATE, DELETE.

## 2.3 La structure de la BDUni

Globalement, chaque version `i` d'un objet est produite à partir d'une réconciliation. Particulièrement, chaque version `N-i` (`N` étant la dernière version) d'un objet pointe vers la réconciliation donnant lieu à la prochaine version `(N-i)+1` de cet objet. Nous avons dérivé un extrait du schéma logique de la BDUni concernant les adresses (voir la Figure 5).

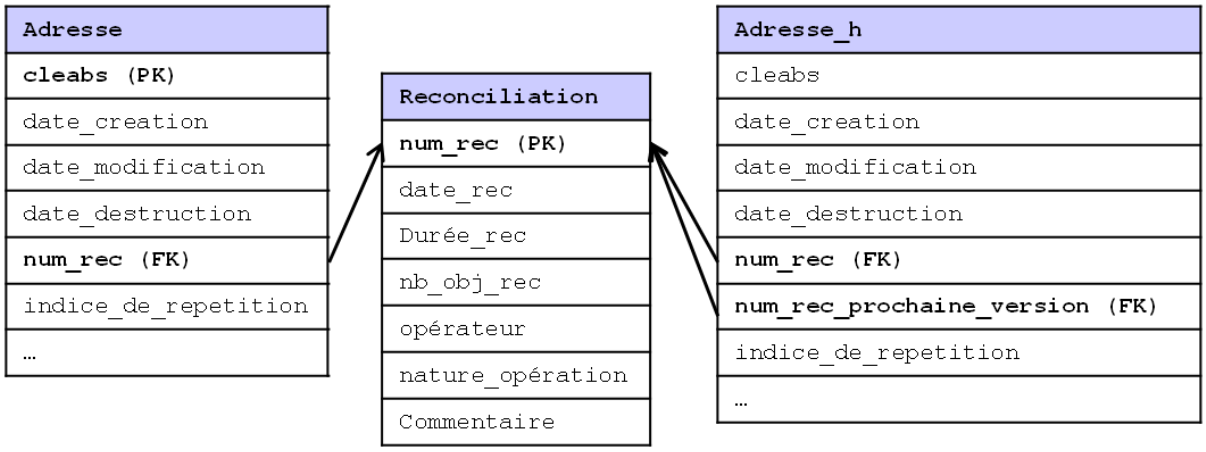



**Figure 5 : Extrait du schéma logique de la BDUni concernant les adresses**

La BDUni contient toutes les dernières versions des objets saisis par les opérateurs. Ces objets, indépendamment de leur type (ex : adresse, bâtiment remarquable, etc.), contiennent les champs suivants : `cleabs`, `date création`, `date modification`, `date destruction`, `numéro réconciliation`, `methode` qui correspondent respectivement à l'identifiant de l'objet dans la base de données (et clé primaire de la table), à la date de création, de dernière modification et de destruction de l'objet (aucun objet n'est jamais supprimé de la base), au numéro du processus de réconciliation lancé par le serveur qui a donné lieu à cette version de l'objet, et à la méthode d'acquisition de l'objet. Par exemple, la dernière version (3ème) d'un objet adresse identifié avec la valeur `"ADRNIVX_0000000287165726"` (`"ADR_2"` pour simplifier) et quelques uns de ses attributs thématiques comme `num voie`, `indice de repetition` et `nom voie`, ont été extraits de la table Adresse et sont présentés ci-dessous :

```
V3 : (cleabs:"ADR_2", date_creation: "2011-08-09 14:20:03",
date_modification:"2012-06-28 16:28:37", date_destruction:"", num_rec:
7795038, méthode:"terrain", num_voie:9, indice_de_repetition:"BIS",
nom_voie:"AV DU BOIS")
```

La BDUni garde aussi toutes les versions des objets qui ont été saisis depuis sa constitution sous la forme d'un historique des données. Chaque version, indépendamment de son type, contient les champs suivants : `cleabs`, `date création`, `date modification`, `date destruction`, `numéro réconciliation`, `numéro réconciliation prochaine version`, `et méthode` qui correspondent respectivement à l'identifiant de l'objet concernant cette version (il ne peut pas être la clé primaire), à la date de création, de dernière modification et de destruction de cette version, au numéro du processus de réconciliation lancé par le serveur qui a donné lieu à cette version de l'objet, au numéro de réconciliation qui donne lieu à

la prochaine version (logiquement cette valeur ne peut qu'être remplie quand une nouvelle version arrivera), et à la méthode d'acquisition de l'objet. Particulièrement, le champ `numéro réconciliation prochaine version` est le lien entre une version de l'objet et le processus de réconciliation qui a donné lieu à sa version précédente (elles partagent évidemment la même clé cleabs). Par exemple, la version actuelle (version #3) et toutes les anciennes versions (version #2, #1 et #0) de l'objet adresse `"ADR_2"` extraites de la table historique des adresses sont listées de la version la plus récente à la version la plus ancienne ci-dessous (la date de modification et les attributs thématiques qui ont changé entre toutes les versions sont soulignés en gras) :

V3 : (cleabs:ADR_2, date_creation:2011-08-09 14:20:03, **date_modification:2012-06-28 16:28:37,** date_destruction:, num_rec:7795038, num_rec_prochaine_version: 8054216, **méthode:terrain,** num_voie:9, **indice_de_repetition:BIS, nom_voie:AV DU BOIS**)

V2 : (cleabs:ADR_2, date_creation:2011-08-09 14:20:03, **date_modification:2011-12-29 14:54:34,** date_destruction:, num_rec:6936925, num_rec_prochaine_version:7795038, **méthode:prélocalisé,** num_voie:9, **indice_de_repetition:BIS, nom_voie:AV DU BOIS**)

V1 : (cleabs:ADR_2, date_creation:2011-08-09 14:20:03, **date_modification:2011-09-15 14:32:11,** date_destruction:, num_rec:6496888, num_rec_prochaine_version:6936925, **méthode:prélocalisé,** num_voie:9, **indice_de_repetition:B, nom_voie:AV DU BOIS**)

VO : (cleabs:ADR_2, date_creation:2011-08-09 14:20:03, **date_modification:2011-08-10 18:47:49,** date_destruction:, num_rec:6333982, num_rec_prochaine_version:6496888, **méthode:prélocalisé,** num_voie:9, **indice_de_repetition:B, nom_voie:**)

Pour résumer chronologiquement, la version 0 correspond à la création de l'objet `"ADR_2"` (son identifiant et sa géométrie) et à l'attribution du champ `indice_de_repetition` à la valeur `"B"`. La version 1 correspond au changement de la valeur du champ `nom_voie` à la valeur `"AV DU BOIS"`. La version 2 correspond à un changement du champ `indice_de_repetition` à la valeur `"BIS"`. La version 3 (actuelle) correspond à un changement du champ `méthode` à la valeur `"terrain"`.

Les informations sur les différents processus de réconciliation (processus décrit dans la sous-section 2.2) sont stockées par le serveur dans la table de réconciliation sous la forme des champs suivants : `numéro réconciliation`, `date réconciliation`, `durée réconciliation`, `nombre d'objets réconciliés`, `opérateur`, `nature operation`, `commentaire` qui correspondent respectivement au numéro du processus de réconciliation donné par le serveur (clé primaire de cette table), la date et la durée du processus, le nombre d'objets impactés par ce processus, l'identifiant de l'opérateur, la nature du changement, et un commentaire additionnel. Par exemple, toutes les informations sur les quatre processus de réconciliation, `#7795038`, `#6936925`, `#6496888`, `#6333982`, qui ont donné lieu aux quatre versions de l'objet adresse `"ADR_2"` sont listées ci-dessous (de la réconciliation la plus actuelle à la plus ancienne) :

**Rec** #7795038 donnant lieu à la version 3 : (num_rec:7795038, date_rec:2012-06-28 16:28:37, duree_rec:11, nb_obj_rec:191, opérateur:TGuillon, nature_opération:Adressage, commentaire:)

**Rec** #6936925 donnant lieu à la version 2 : (num_rec:6936925, date_rec:2011-12-29 14:54:34, duree_rec:21, nb_obj_rec:124, opérateur:EDorange-Pattoret, nature_opération:Adressage, commentaire:Ecriture en toutes lettres indice de repetition)

**Rec** #6496888 donnant lieu à la version 1 : (num_rec:6496888, date_rec:2011-09-15 14:32:11, duree_rec:5, nb_obj_rec:4, opérateur:TGuillon, nature_opération:Adressage, commentaire:)

**Rec** #6333982 donnant lieu à la version 0 : (num_rec:6333982, date_rec:2011-08-10 18:47:49, duree_rec:8, nb_obj_rec:200, operateur:TGuillon, nature_operation:correction, commentaire:"")

Nous constatons que pour connaître ce qui a changé entre les différentes versions, il faut recourir systématiquement à des opérations de comparaison des valeurs d'attributs entre deux versions de l'objet. Le champ `commentaire` en texte libre peut être aussi précieux pour comprendre la nature de l'évolution, ainsi les commentaires des opérateurs autour des réconciliations précédentes (s'ils sont remplis) nous permettent d'interpréter ce qui a changé. Par exemple, le commentaire rempli par l'opérateur entre la version #1 et #2 est `"Ecriture en toutes lettres indice de repetition"`. De cette manière, un humain peut savoir que la valeur de l'attribut `indice de repetition` a changé.

## 2.4 Le projet Échanges

La mise à jour du RGE® s'appuie sur des partenariats avec des administrations départementales (ex : les conseils généraux), nationales (ex : Le Cadastre) et régionales, ainsi qu'avec des sociétés possédées par des fonds publics (ex : la Poste et) qui mettent à jour eux-mêmes des bases de données indépendantes de l'IGN dont le contenu est partiellement commun avec celui du RGE®. Le bénéfice principal de cette démarche partenariale, est de réduire les couts de la mise à jour des données pour l'IGN, et pour les partenaires, des droits d'usage et des mises à jour du (ou d'une partie du) RGE®. Afin de faciliter l'incorporation des mises à jour dans les bases de données partenaires, elles sont livrées (en XML) sous la forme de différentiels de données géographiques grâce au modèle de diffusion des évolutions proposé par Badard (2000) (thèse effectuée au COGIT). Ces démarches sont en cours dans le cadre du projet Echanges (Viglino 2010; Viglino 2011). La plate-forme Web RiPart[5] (Remontée d'Information Partagée) développée sur la base de l'API Géoportail[6] de l'IGN et mise en place dans le cadre du projet Echanges facilite la gestion des remarques sur les données et des informations transmises par les partenaires sur le Web (Viglino 2011). Le partenaire possède un compte utilisateur dans RiPart. Il peut signaler une remarque en transférant un fichier joint (PDF, JPG, DOC), une trace GNSS (en GPX ou KML) permettant de visualiser un tracé géoréférencé, ou de dessiner un croquis sur un fond Géoportail avec les outils disponibles sous RIPart. Ces documents doivent être accompagnés d'une description en texte libre précisant la remarque. Le partenaire est invité à adhérer à des groupes génériques d'utilisateurs. Chacun de ces groupes s'intéresse à un thème en particulier comme par exemple la randonnée. Cela permet de définir le cadre lequel les remarques sont formulées pour mieux cerner le profil de l'initiateur d'une remarque afin de faciliter la qualification des remontées (Viglino 2011). D'autre

[5] http://ripart.ign.fr/
[6] http://api.ign.fr

part, RiPart rattache les utilisateurs qui participent régulièrement à la remontée d'informations à un groupe d'utilisateurs reconnus, cela permettrait de noter que le système accorde du crédit à ses remarques. La Figure 6 montre un exemple d'une remarque dessinée par un partenaire pour signaler la présence d'un nouveau parking. La remarque est ensuite transférée vers le serveur afin d'être validée par un opérateur de la MAJEC. Si la remarque a été réalisée par un auteur dans le cadre de son appartenance à un groupe, la remarque est envoyée à l'opérateur de la MAJEC concerné par le thème du groupe. Ensuite, les objets correspondants sont éventuellement saisis (créés, modifiés, ou supprimés) dans la BDUni par l'opérateur, et le partenaire reçoit une réponse expliquant si sa remarque a bien été prise en compte ou non.

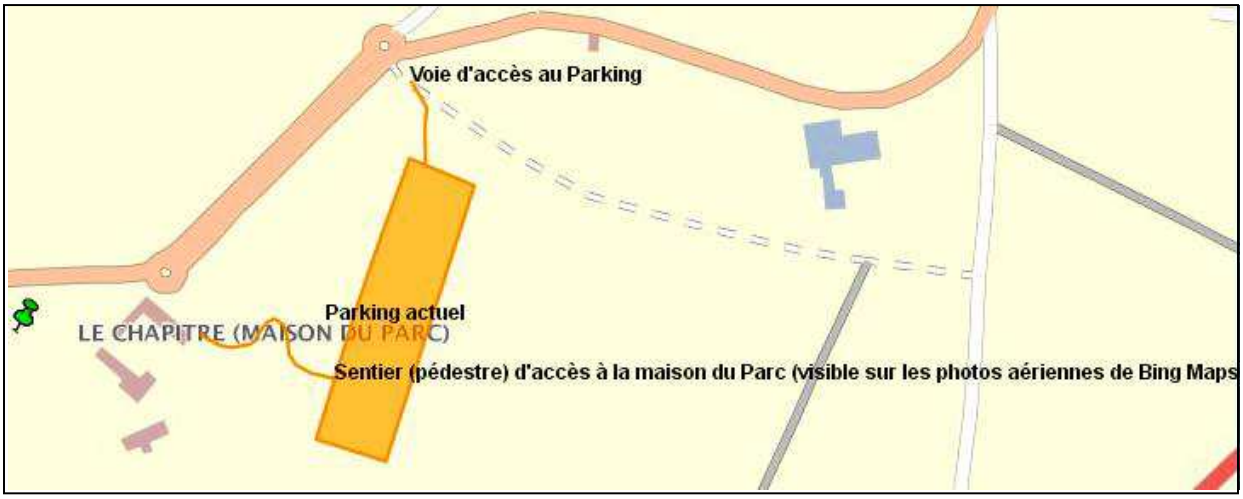



**Figure 6 : L'emplacement d'un nouveau parking est dessiné sur la carte grâce à RiPart**

RiPart est aussi disponible pour les partenaires via un API REST (Fielding 2000) où les remarques sont accessibles sur le Web à partir d'un URL. L'API permet (via une requête HTTP) de soumettre une nouvelle remarque, et de lister les remarques effectuées par un partenaire avec la réponse obtenue par l'opérateur. Une remarque comporte les informations suivantes : un numéro unique, la date de postage, l'auteur, son appartenance à un groupe, les coordonnées géographiques, les produits IGN concernés, une description et un code sur l'avancement de la remarque. Ces codes sont limités aux valeurs suivantes : « reçue dans nos services », « en cours de traitement », « prise en compte », « rejetée ». Le code XML présenté ci-dessous montre un extrait de la réponse de l'API concernant une remarque identifiée avec `305`. Cette remarque a été réalisée par un partenaire identifié comme `darrepac` signalant l'absence d'un sentier sur le produit IGN carte en papier Top25. Elle est suivie d'une réponse de la part de `vautard` (IGN 2011a). Par ailleurs, grâce à l'expérience acquise avec RiPart, l'IGN vient de mettre en ligne sur le site Géoportail IGN, un outil de signalement d'erreurs, d'évolutions ou d'omissions, afin de permettre au grand public de participer à la mise à jour du RGE®.

```
<!-- Remarque #305 ->
<ID_GEOREM>305</ID_GEOREM>
<DATE>2010-09-29 15:40:42</DATE>
<MAJ>2010-10-01 13:26:01</MAJ>
<DATE_VALID>2010-10-01 13:26:01</DATE_VALID>
<LON>6.83209712586289</LON>
<LAT>44.7545129208894</LAT>
```

```
<STATUT>valid</STATUT>
<DEPARTEMENT>Hautes Alpes</DEPARTEMENT>
<COMMUNE>Molines-en-Queyras</COMMUNE>
<COMMENTAIRE>
Le sentier reliant l'oratoire et la route à l'Est et qui passe par le point
noté est manquant sur la carte IGN top25. Ce sentier est évident sur la
BDOrtho et aussi dans la réalité sur le terrain (il y a même un banc installe
en dur au milieu afin que les gens profitent du panorama) On voit d'ailleurs
que le chemin se poursuit après la route a l'Est et que cette poursuite est
bien notée sur la carte top25.
</COMMENTAIRE>
<AUTEUR>darrepac</AUTEUR>
<!-- Réponse #149 ->
<GEOREP>
<ID_GEOREP>149</ID_GEOREP>
<AUTEUR>vautard</AUTEUR>
<STATUT>valid</STATUT>
<DATE>2010-10-01 13:25:13</DATE>
<REPONSE>
Bonjour,
Après vérification, un sentier existe bien dans nos bases de données à
l'endroit de votre remarque. Vous pouvez d'ailleurs le visualiser en
affichant la couche BDTopo et en zoomant suffisamment. Il devrait donc
figurer sur la prochaine édition de la carte. Bien cordialement,
</REPONSE>
</GEOREP>
```

## 2.5 Le contrôle de qualité des données

Après la collecte et l'édition des données dans la BDUni, les contrôles de qualité sont effectués par l'unité de qualité en suivant les spécifications qui précisent les seuils acceptables pour les indicateurs de qualité (voir section I.1 pour rappel). Pour connaître la qualité d'un jeu de données géographiques, il s'agit de le comparer au terrain nominal et de mesurer les écarts constatés (IGN 2009b). Avec deux opérateurs, toute l'activité d'évaluation dure environ trois mois. L'unité minimale d'évaluation est le département et le contrôle est effectué par sondages exhaustifs sur un échantillon composé de quelques communes du département concerné, choisies pour leur représentativité du paysage (urbain/rural, plaine/montagne…). Pour certains types d'objets jugés importants comme le réseau routier, le réseau ferré, le réseau électrique, et les bâtiments publics, l'échantillon peut être le département complet (IGN 2011b). Les types d'évaluations de qualité effectuées sont les contrôles thématiques mesurant l'exhaustivité et la précision sémantique, et les contrôles géométriques mesurant la précision géométrique. Dans tous les cas, l'opérateur comptabilise les différences entre le contenu de la base de données et le terrain vu au travers des spécifications à la date de la prise de vue aérienne. Une différence est soit un déficit correspondant à un objet manquant dans la base, soit un excédent correspondant à un objet en trop dans la base. Une confusion correspondant à un objet présent dans la base mais dont les attributs sont mal codés (IGN 2009b). Par exemple, une erreur d'emplacement, appelée une confusion, d'une mairie à Fresney-le-Vieux dans le département de Calvados (montrée dans la Figure 7) est détectée lors d'une procédure de contrôle.

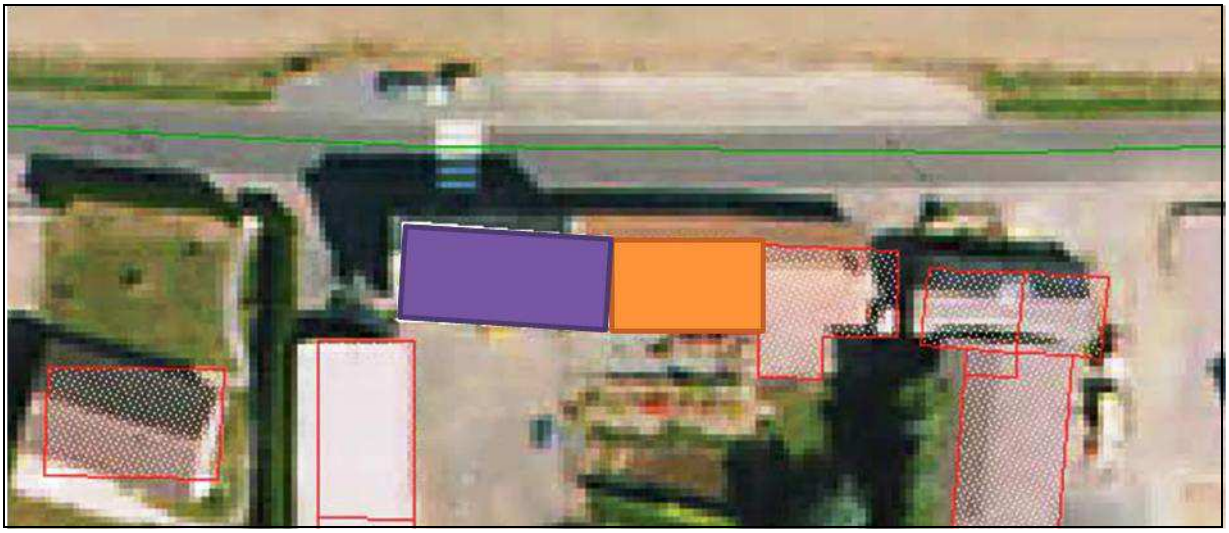


**Figure 7 : Erreur d'emplacement d'une marie à Fresney-le-Vieux dans le département Calvados, en violet la marie dans la BDUni et en orange la position de la mairie sur le terrain (IGN 2009a)**

# 3 La production communautaire d'un contenu géographique et la gestion de sa qualité

Cette section décrit la production collaborative d'un contenu géographique du point de vue d'un projet communautaire de cartographie collaborative et la façon dont la qualité d'un tel type de contenu est gérée. D'abord, nous présentons d'une manière générale une caractérisation de certains contenus géographiques communautaires actuellement disponibles en ligne. Puis, nous nous concentrons sur le projet communautaire OSM. Dans le cadre de ce travail de thèse, nous avons proposé début 2012 sur la liste de diffusion OSM France un sondage en ligne aux contributeurs de cette communauté, duquel 58 contributeurs ont participé en exprimant leurs avis et en proposant des améliorations sur certains aspects du projet OSM. Nous avons également participé le 21 janvier 2012 à une activité OSM de cartographie des éléments d'accessibilité pour personnes en fauteuil roulant dans le centre-ville de la ville de Montpellier organisée par l'association Montpel'libre[7]. Cette activité et les réponses au sondage (voir l'annexe A) nous ont permis de connaître (en dehors de la documentation officielle) les particularités de la production d'un contenu géographique par une communauté comme OSM.

## 3.1 Vers une caractérisation des contenus géographiques communautaires

Le phénomène où les communautés en ligne deviennent des producteurs de données géographiques a été baptisé par Goodchild (2007) l'Information Géographique Volontaire (plus connu par son acronyme en anglais VGI pour *Volunteered Geographic Information*), ou encore néogéographie (Turner 2006). Le projet OSM avec d'autres projets comme Google Map Maker (GMM)[8] et Wikimapia[9] sont des exemples de projets produisant du VGI. Ces projets visent à

[7] http://montpel-libre.fr/
[8] http://www.google.fr/mapmaker
[9] http://www.wikimapia.org/

cartographier toutes les entités géographiques du monde. Ces projets semblent « généralistes » par rapport au contenu qu'ils produisent, dans le sens où il n'y a pas d'application spécifique *a priori*. Par ailleurs, l'aspect géographique de la donnée produite dans ce type de projets est complexe car les formes, les localisations, et certaines topologies des objets sont saisies. Dans ces projets, le modèle de données a tendance à changer très fréquemment, notamment pour OSM où, dès qu'un nouveau besoin arrive, le modèle est enrichi par quelques contributeurs pour y répondre. Par exemple, la cartographie des circonscriptions électorales pendant les élections législatives françaises 2012[10] était un besoin requis par une organisation non-gouvernementale (ONG)[11]. Pour GMM et Wikimapia, le modèle est changé par les administrateurs du projet.

D'autres exemples de VGI sont les projets SpotCrime[12] et Ushahidi[13] dont nous considérons les contenus comme « non-généralistes ». Ils sont considérés ainsi car ils existent pour remplir un besoin spécifique, comme la cartographie de la criminalité pour SpotCrime et la cartographie sociale pour Ushahidi. Le but est donc de saisir les objets correspondant à ce besoin. Dans ces projets non-généralistes, le modèle de données est prédéfini. Dans la suite, nous nous concentrerons sur OSM, un projet né en 2004 à Londres. Aujourd'hui, il connait un succès important en termes de nombre de contributeurs (1000 contributeurs actifs[14]), de réactivité (visible dans les listes de discussion), de documentation en ligne (Wiki OSM) et de volume de contenu produit et diffusé librement[15] en ligne.

## 3.2 La représentation des entités géographiques dans OSM

OSM définit quatre éléments fondamentaux (ou « primitifs ») pour la représentation des entités géographiques, comme les nœuds, les chemins, les relations, et les tags. Une entité géographique est représentée soit sous la forme d'un nœud, soit sous la forme d'un chemin. Les géométries ponctuelles (ex : une bouche de métro) sont représentées par un nœud. Les géométries polylinéaires ouvertes (ex : une piste de ski), fermées (ex : un carrefour giratoire) ainsi que les géométries surfaciques (ex : un lac) sont représentées par un chemin. Il faut préciser que toutes les géométries sauf les ponctuelles sont d'autres types de chemins. Un élément de regroupement d'objets est une relation[16]. Les polygones avec des trous, comme un bâtiment avec une cour ou une forêt avec une clairière, doivent être construits avec l'aide des relations (Ramm et al. 2010). Les relations visent à créer un lien de proximité entre les objets, par exemple « cette entrée amène à cette station de métro », « il n'est pas possible de tourner de cette rue à cette rue » (OSM 2012f). Un autre exemple est la représentation de l'itinéraire cyclable « La Loire à vélo » dans OSM, un itinéraire cyclable tout au long du fleuve traversant des avenues, des ponts, des quais, des places, et des tronçons d'un itinéraire de ferry. Cet itinéraire a été représenté comme une relation OSM identifiée avec la valeur `31297` et

[10] http://wiki.openstreetmap.org/wiki/FR:Circonscription_législative

[11] Voir la demande de l'ONG Regards Citoyens dans sur la liste de diffusion : http://lists.openstreetmap.org/pipermail/talk-fr/2012-May/043607.html

[12] http://www.spotcrime.com/

[13] http://ushahidi.com/

[14] Émission Le dessous des cartes diffusée le 12 juin 2012 sur ARTE

[15] Données sous licence CC-By-Sa : toute réutilisation doit respecter ces deux critères, citation de la source, et partage des modifications sous les mêmes termes du contrat de licence

[16] Le mot relation est réservé dans ce travail de thèse à l'instanciation d'un type de relation entre deux types d'objet, dans le sens classique de modélisation de données géographiques.

contenant les références vers l'identifiant des objets OSM représentant les ponts, avenues, etc. correspondants[17] (voir la Figure 8).


**Figure 8 : la relation OSM représentant l'itinéraire cyclable « La Loire à vélo » visualisée sur OSM**

Une propriété descriptive d'une entité est représentée sous la forme d'un tag, cette étiquette, constituée d'une *clé* et d'une *valeur*, est ajoutée à un nœud, un chemin, ou une relation, et est écrite en texte libre par le contributeur de la manière suivante : `clé=valeur`. La clé décrit l'objet d'une manière générale. Par exemple, les routes, les cours d'eau, les voies ferrées doivent avoir, respectivement, les clés `highway`, `waterway`, `railway`. La valeur décrit plus spécifiquement l'objet à partir de la clé. Par exemple, les chemins de montagne, les canaux, et les rails de métro doivent avoir, respectivement, les valeurs suivantes (avec leurs clés) : `highway=track`, `waterway=canal`, et `railway=subway`. Globalement, les tags sont classifiés en trois groupes, ceux visant à décrire un objet matériel du monde réel (ex : les routes, les barrières, les voies cyclables, les monuments historiques, les commerces), d'autres visant à décrire un objet immatériel (ex : les frontières, les itinéraires), et ceux communs à tous les objets (ex : la dénomination, la source). Cette classification a été construite par la communauté OSM et est détaillée sur la page Web Map Features[18]. Néanmoins, cette liste n'est pas exhaustive et l'utilisation de ces tags n'est pas obligatoire, les contributeurs sont encouragés à la suivre afin de garantir une certaine homogénéité des données. Les contributeurs sont également encouragés à utiliser l'outil Web Taginfo[19], un système qui agrège les tags et produit des statistiques sur la fréquence d'utilisation des tags des objets de la base de données OSM. Par exemple, la Figure 9 montre les tags les plus utilisés par les objets OSM (nœuds, chemins) et plus particulièrement par les objets de type chemin. Taginfo permet à un contributeur de trouver par un mot-clé un tag non documenté dans la page OSM des Map Features déjà utilisé par d'autres contributeurs. Des nouveaux tags peuvent être proposés par les contributeurs après une discussion sur la liste de diffusion OSM et en les proposant sur la page Web Proposed Features[20]. Dans le sondage que nous avons proposé, les contributeurs

[17] http://www.openstreetmap.org/browse/relation/31297
[18] http://wiki.openstreetmap.org/wiki/FR:Map_Features
[19] http://taginfo.openstreetmap.org
[20] http://wiki.openstreetmap.org/wiki/Proposed_features

paraissent assez satisfaits du système de tags OSM du fait de sa flexibilité. Néanmoins, plusieurs contributeurs relèvent quelques difficultés :

- recherche de tags non-aisée correspondant à une entité particulière,
- saisie des informations fonctionnelles sans utiliser les tags,
- manque de détail pour les tags proposés à l'heure actuelle pour certains thèmes comme les zones humides,
- lien entre la documentation des tags sur le Wiki OSM, et
- des outils d'édition afin de tagger plus efficacement.
-

| Tag | Objects | | Ways | |
|---|---|---|---|---|
| building=yes | 58 606 698 | 3.54% | 58 390 527 | 40.97% |
| highway=residential | 23 252 843 | 1.40% | 23 250 361 | 16.31% |
| wall=no | 6 841 174 | 0.41% | 6 840 996 | 4.80% |
| highway=service | 6 463 497 | 0.39% | 6 461 934 | 4.53% |
| waterway=stream | 5 605 179 | 0.34% | 5 600 036 | 3.93% |
| highway=unclassified | 5 451 382 | 0.33% | 5 450 750 | 3.82% |
| source=3dShapes | 4 686 207 | 0.28% | 4 631 516 | 3.25% |
| highway=track | 4 551 672 | 0.27% | 4 550 657 | 3.19% |

**Figure 9 : les tags les plus utilisés par les données OSM selon l'outil Web Taginfo**

Chaque objet OSM peut être consulté via l'API REST OSM[21] en fournissant une URL de la forme `http://www.openstreetmap.org/api/0.6/<Typed'objet>/<id_objet>`. La réponse est l'objet en format OSM XML, un format d'échange basé sur XML particulier à OSM[22]. Par exemple, l'objet de type chemin correspondant à un des bâtiments de l'IGN identifié comme `145052693` contenant sept nœuds et trois tags `construction`, `landuse`, `source`, créé le 25 janvier 2012 est présenté ci-dessous[23] :

```
<osm version="0.6" generator="OpenStreetMap server">
      <way id="145052693" visible="true" timestamp="2012-01-
      25T23:36:10Z" version="2" changeset="10498790" user="Pieren"
      uid="17286">
            <nd ref="1503865153"/>
            …
            <tag k="construction" v="yes"/>
            <tag k="landuse" v="construction"/>
            <tag k="source" v="knowledge"/>
      </way>
</osm>
```

[21] Actuellement en version 0.6 : http://wiki.openstreetmap.org/wiki/API_v0.6

[22] Une tentative de définir le schéma DTD du format OSM XML est publiée ici : http://wiki.openstreetmap.org/wiki/OSM_XML/DTD

[23] Requête HTTP : http://www.openstreetmap.org/api/0.6/way/145052693.

A titre d'exemple, la géométrie du nœud identifié `1503865153` qui compose le chemin identifié `145052693` correspondant à un bâtiment de l'IGN, est présentée ci-dessous :

```
<osm version="0.6" generator="OpenStreetMap server">
      <node id="1503865153" version="2" changeset="10356427" lat="48.8437702"
      lon="2.4232988" user="Pieren" uid="17286" visible="true"
      timestamp="2012-01-10T22:50:48Z"/>
</osm>
```

La Figure 10 montre le bâtiment en entier visualisé sur OSM.


**Figure 10 : le chemin OSM représentant le bâtiment de l'IGN à Saint-Mandé visualisé sur OSM**

## 3.3 Le processus de contribution à OSM

Les contributeurs suivent plusieurs méthodes de contribution pour l'acquisition de la nouvelle donnée sur le terrain, à l'aide d'un récepteur GNSS ou par le tracé des images satellites. Ensuite, ils incorporent la nouvelle donnée ainsi que d'autres sources externes en éditant le contenu de la base de données centrale OSM en utilisant les outils d'édition mis en place par la communauté de développeurs. Dans tous les cas, les contributeurs sont encouragés à toujours citer la source de la donnée. Plusieurs valeurs sont prédéfinies : `images satellites`, `photo personnelle`, `connaissance locale`, `connaissance commune`, `dictaphone`. Le client d'édition en ligne Potlatch[24] est conseillé aux contributeurs débutants et le client lourd hors-ligne JOSM[25] est conseillé aux contributeurs avec une longue expérience dans le projet. Les modes principaux de contribution à OSM, selon notre sondage, sont listés au Tableau 1.

| **Mode de contribution à OSM** | **Nombre et pourcentage de contributeurs utilisant ce mode de contribution** |
|---|---|
| Trace des images satellites fournies par Yahoo! Maps et Bing Maps | 56 (96,6 %) |
| Chargement de sources de données autoritaires disponibles (ex : cadastre, Corine Land Cover) | 41 (70,7 %) |
| Collecte de traces GNSS et téléchargement sur le serveur OpenStreetMap | 39 (67,2 %) |

**Tableau 1 : Les trois principaux modes de contribution selon un sondage proposé aux contributeurs OSM France**

[24] http://wiki.openstreetmap.org/wiki/Potlatch_2
[25] http://josm.openstreetmap.de/

Le tracé des nouvelles géométries sur des images satellites légalement fournies par les sociétés Yahoo! et Microsoft est supporté par les outils principaux d'édition OSM. Ce mode d'acquisition est conseillé pour l'acquisition de certains objets comme des aires forestières, des cours d'eau (lacs et rivières) ainsi que certaines voies ferrées (Ramm et al. 2010). Les formes des bâtiments et des réseaux routiers peuvent également être acquises, néanmoins les localisations ne sont pas toujours correctes. L'actualité de l'image satellite peut être aussi un problème car certaines de ces images sont anciennes. Avant la sortie sur le terrain, Ramm et al. (2010) conseillent le tracé des images satellites sur des zones où il n'y a aucune donnée afin de construire un « squelette » de la zone, puis sur place, de remplir les tags des objets correspondants et ainsi vérifier la donnée tracée.

Un autre mode de contribution est l'import massif de sources de données externes. Les imports doivent être auparavant planifiés et discutés sur la liste de diffusion et la licence de la source de données doit être vérifiée avec prudence. Depuis 2008, les plans vectorisés du cadastre en France sont disponibles en ligne au format PDF (image en SVG) par un service Web élaboré par la Direction Générale des Finances publiques (DGFiP). Il a donc été possible d'obtenir le cadastre, les bâtiments, les plans d'eau, ainsi que certains cours d'eau et monuments. Un post-traitement semi-automatique de ces données a été nécessaire comme le retrait de nœuds dupliqués, la correction des bâtiments découpés ou encore le calcul des incohérences de superposition entre les bâtiments OSM et le cadastre (OSM 2012g). Néanmoins, la qualité des données obtenues peut être très faible malgré les traitements. Depuis 2009, la base de données européenne d'occupation biophysique des sols Corine Land Cover (CLC) est également mise à disposition en ligne (par téléchargement et par service Web) en France pour OSM par le Service de l'Observation et des Statistiques du Commissariat Général au Développement Durable (CGDD) du Ministère de l'écologie (OSM 2012h). Certains polygones ont été automatiquement importés dans OSM, et d'autres sont importés manuellement au fur et au mesure par la communauté à l'aide des outils de visualisation de l'état de l'import (voir la Figure 11).

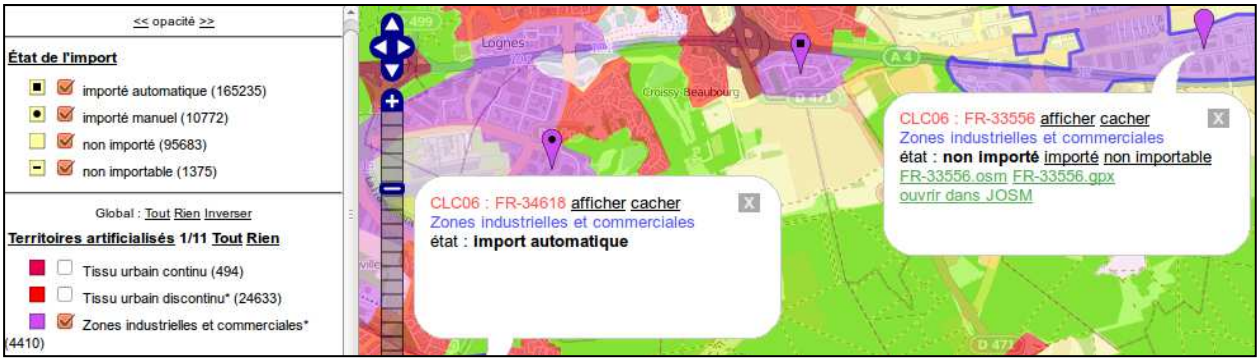



**Figure 11 : deux zones CLC : l'une importée automatiquement et l'autre pas encore importée dans OSM, visualisées sur un outil de visualisation de l'état de l'import de données CLC**[26]

Plus récemment, les données publiques[27] (connu comme l'*open data*) sont en train d'être importées dans OSM. Pour ceci, l'interface Web de visualisation PlaceMaker[28] (voir la Figure

[26] http://clc.openstreetmap.fr/cgi-bin/index.py
[27] http://www.data.gouv.fr/

12) est mise à disposition par la communauté de développeurs OSM afin de visualiser les données, les éditer, puis les importer dans OSM (OSM 2012c). A l'heure actuelle, les points de contact du réseau postal français et la base de données des limites administrative GEOFLA sont disponibles respectivement grâce à La Poste et à l'IGN. Des modules (*plugins*) dédiés à la lecture des données publiques (en format Excel, GML, etc.) sont maintenant disponibles dans JOSM afin de faciliter l'import de telles données dans OSM.

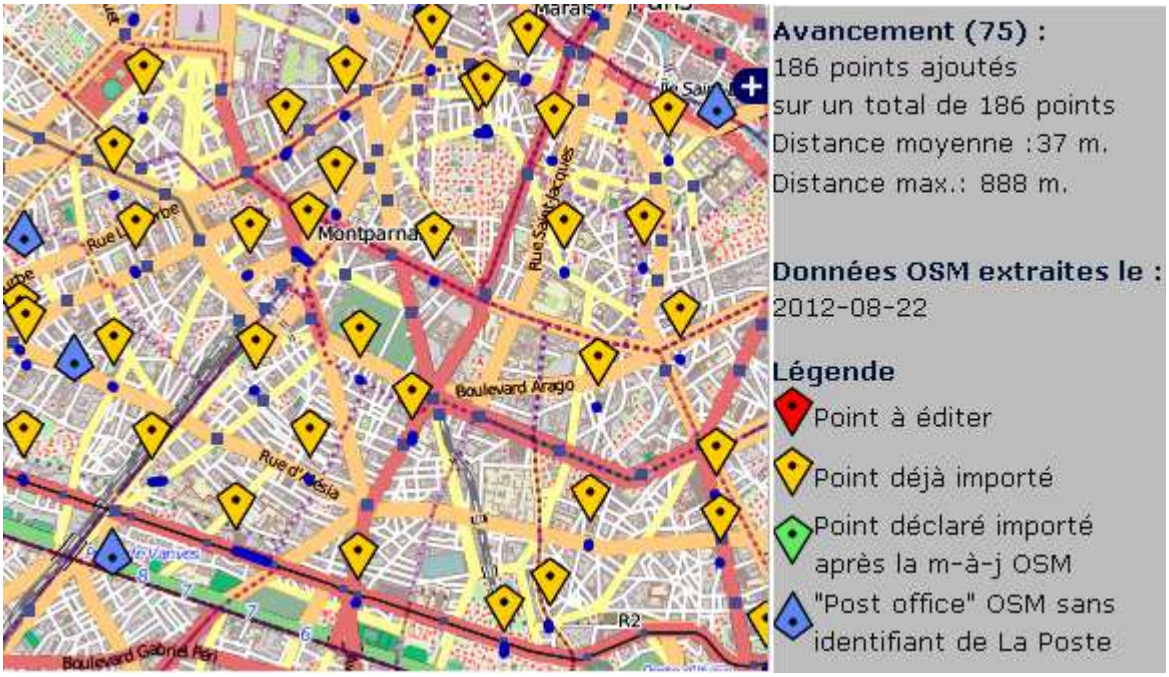



**Figure 12 : une carte en ligne Place Maker est mise à disposition par un contributeur pour faciliter l'incorporation des bureaux de postes (en open data) dans OSM**

L'acquisition de données sur le terrain est souvent effectuée dans le cadre d'une activité particulière à OSM entre les contributeurs, c'est ce qu'on appelle une cartopartie. Elle est régulièrement organisée par la communauté afin de cartographier une zone particulière entre de nombreux contributeurs. A côté des listes de diffusion, celles-ci restent comme un moyen privilégié de communication entre les contributeurs. Une cartopartie peut aussi être proposée par une organisation non gouvernementale (ONG), comme par exemple la cartopartie du 21 janvier 2012 organisée par l'association Montpel'libre[29] à Montpellier. Elle a compté sur la participation de bénévoles qui n'avaient pas nécessairement de connaissances d'OSM et de contributeurs avec une longue expérience dans OSM qui ont guidé l'acquisition. L'objectif était d'enrichir la donnée OSM en ajoutant des nouveaux tags dédiés à l'accessibilité des personnes en fauteuil roulant, définis auparavant par l'acquisition faite par les organisateurs[30]. Ces tags permettent d'indiquer la présence de rampes, de superficies en pierre, etc. Dans une cartopartie, l'organisateur découpe auparavant en plusieurs parties la zone à sonder, selon les

[28] http://osm.vdct.free.fr/postes/index.html
[29] Voir la vidéo ici : http://www.dailymotion.com/video/xoh2sj_montpellier-une-cartopartie_news
[30] Http://wiki.openstreetmap.org/wiki/FR:Wheelchair_routing

nombre de participants, et chaque partie est imprimée en format papier en utilisant le service OSM Walking papers[31]. L'organisateur constitue plusieurs équipes, chacune composée d'un contributeur avec une longue expérience dans OSM et des bénévoles intéressés à l'activité de l'ONG. Chaque équipe est en possession des copies de cartes en papier de la partie correspondante (la Figure 13 à g.) et un formulaire papier (la Figure 13 à d.) afin de noter les objets pour lesquels des informations sur leur accessibilité sont disponibles. Une fois la cartopartie finalisée, les cartes en papier avec les annotations des participants sont rassemblées par les organiseurs qui font la saisie ultérieurement sur un client d'édition des données OSM. Ces cartes en papier annotées peuvent également être scannées en fichiers image et téléchargées sur le serveur OSM afin de servir comme un support documentant la collecte. Le contributeur peut également fournir des fichiers sons MP3 et des photos géolocalisées acquis durant la collecte.

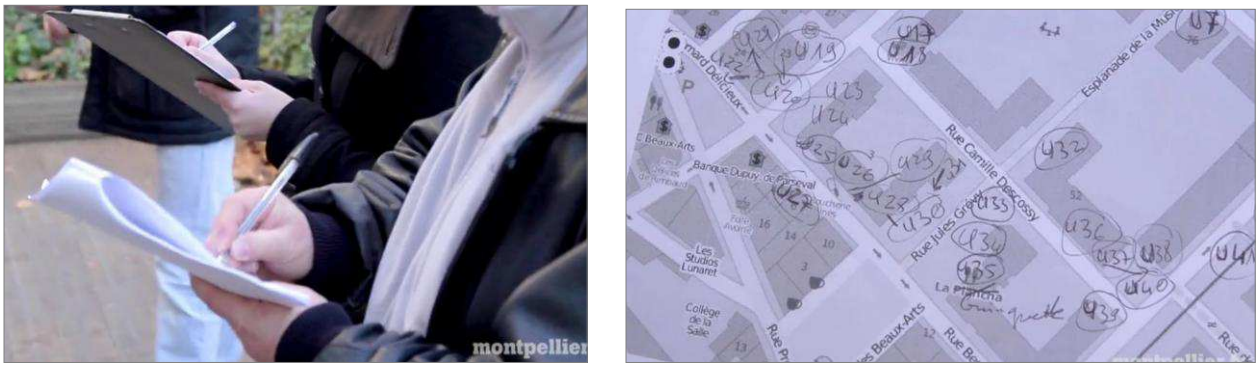



**Figure 13 : (g.) L'utilisation des cartes papiers et (d.) des formulaires papiers lors de la cartopartie du 21 janvier 2012 organisée par l'association Montpel'libre à Montpellier**

L'acquisition des données sur le terrain se fait également à l'aide d'un récepteur d'antenne GNSS (dans le cadre d'une cartopartie ou non). Ce mode était la manière de contribuer à l'origine du projet, utilisé principalement pour collecter les géométries initiales des réseaux routiers, cyclistes et piétons, aujourd'hui ce mode reste toujours très populaire selon notre sondage. Il est également possible de se servir des logiciels développés sur *smartphones* ou tablettes[32] pendant l'acquisition, comme par exemple, de prendre des photos géolocalisées ou de renseigner des nouveaux points d'intérêt avec iLOE pour IPhone ou JPSTrack pour Android. Afin de préparer l'acquisition à l'aide d'un récepteur antenne GNSS, le contributeur choisit une zone à changer, par exemple, un nouveau quartier. Ensuite, il effectue un parcours exhaustif à pied, en voiture, ou au vélo sur la zone choisie et enregistre ses traces à l'aide de l'appareil GNSS. L'utilisation de certains récepteurs d'antennes GNSS avec peu de précision métrique (autour de 5 mètres) est une faiblesse d'OSM par rapport aux méthodes de collecte traditionnelles. Néanmoins, Ramm et al. (2010) indiquent que ceci ne pose pas un problème car

[31] Walking papers (http://walking-papers.org) est un service qui permet d'imprimer un extrait de la carte OSM, de l'annoter, de scanner vos annotations en retour de manière à ajouter vos nouvelles données à OSM

[32] Nombreuses applications mobiles existent pour OSM, version IPhone : http://wiki.openstreetmap.org/wiki/Apple_iOS et Android : http://wiki.openstreetmap.org/wiki/Android

le but d'OSM n'est pas de produire des données à grande échelle avec des limites très précises, mais de fournir des données pour des aires de taille moyenne et des villes. Par contre, bien indiquer la topologie des objets peut être plus importante que la précision géométrique (principalement pour le réseau routier). Suite à l'acquisition sur le terrain, les traces GNSS récoltées (souvent en format GPX) sont incorporées dans la base de données par un client d'édition de données OSM, et peuvent même être stockées en GPX dans OSM comme un support documentant la collecte.

Suite à l'acquisition des traces GNSS, le contributeur les incorpore en éditant le contenu de la base de données correspondant à la zone sondée. Il démarre une session sur un client OSM, par exemple JOSM, en fournissant son identifiant utilisateur et son mot de passe. Le contributeur charge les traces GNSS sur le client qui peut donc calculer la zone à changer ; sinon le contributeur peut toujours sélectionner la zone lui-même en dessinant un rectangle de la zone à changer sur un fond carte OSM. Ensuite, le serveur transfère les données OSM correspondant à la zone sondée du serveur vers le client. De cette façon, le contributeur peut confronter ses traces GNSS avec les données centrales. La Figure 14 (à g.) montre sur JOSM l'absence dans les données OSM d'un tronçon de route sur le Boulevard des Cents Arpents (département Seine-et-Marne) dont la géométrie est mise en évidence pour une trace GNSS (en bleu)[33]. Le contributeur dessine alors les objets manquants dans OSM mis en évidence par les traces GNSS (voir la Figure 14 à d.) en considérant, pour cet exemple, la consigne de création d'un nœud de jonction pour les tronçons de routes qui s'intersectent (OSM 2012b).

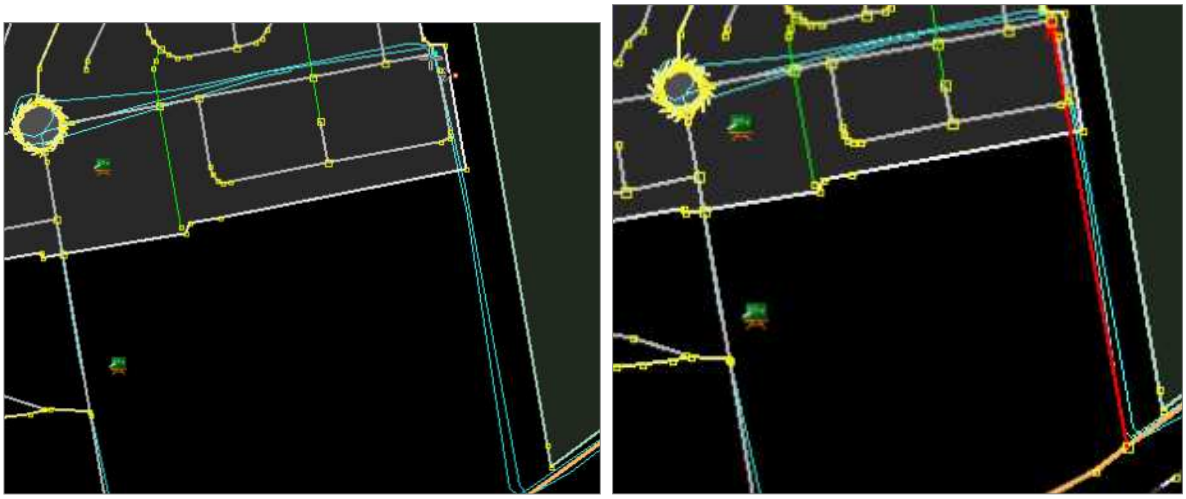


**Figure 14 : (g.) Trace GNSS en bleue correspondante à un tronçon de route du Boulevard des Cents Arpents et, (d.) tracé du nouveau tronçon (en rouge) cohérent à la trace**

Suite à la saisie de la zone concernée en utilisant un client d'édition OSM, le contributeur déclenche le transfert de ses données vers le serveur (cas du client lourd comme JOSM). Avant le transfert, le client montre un recensement de tous les changements, exécute des procédures locales afin de détecter des incohérences dans les données, et les signale au contributeur (voir la Figure 15) comme des erreurs. Par exemple, la présence de nœuds ou de chemins dupliqués, d'une ligne côtière sans une étendue de terre d'un côté, de fautes de frappe des clés et valeurs, engendre des avertissements, tout comme des routes superposées, des routes se

[33] Traces GPS collectées le 5 décembre 2010, voir la vidéo entière de la saisie sur : http://www.youtube.com/watch?v=rv3a_HnMtBw

croisant entre elles ou des chemins sans tags. Le contributeur corrige les erreurs (pas nécessairement les avertissements) et confirme le transfert des données vers le serveur OSM.

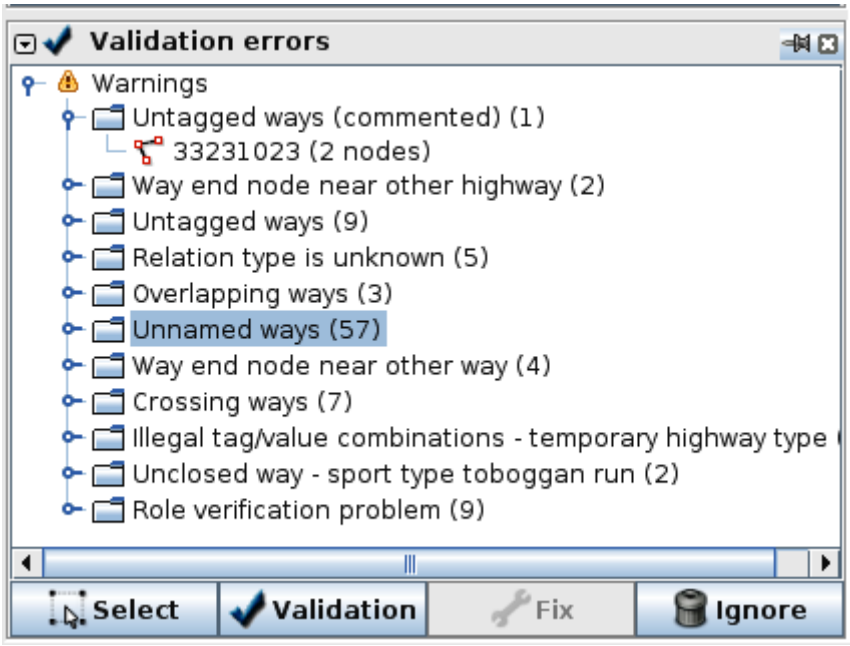



**Figure 15 : Erreurs détectées lors de l'exécution de procédures locales de validation des données dans JOSM**[34]

Dans le cas de conflits de versions, par exemple un nœud supprimé par le contributeur est utilisé par un chemin dans le serveur, ou un objet modifié par le contributeur qui a été supprimé par un autre contributeur, les changements ne sont pas effectués dans la base de données (Ramm et al. 2010). Le serveur envoie plutôt l'information sur les conflits au contributeur qui doit les résoudre à l'aide d'une interface graphique (ex : le plugin JOSM pour la résolution de conflits)[35]. Cet outil aide le contributeur à visualiser les divergences entre deux versions afin qu'il puisse choisir la meilleure version de l'objet ou fusionner les tags de sa version et celle du serveur pour comparer par exemple des versions différentes entre le client et le serveur. Enfin, le contributeur doit renvoyer ses données en résolvant si possible, les conflits.

## 3.4 Les messages échangés client/serveur

Les nouvelles versions d'un objet proposées par un contributeur pendant une session d'édition de contenu OSM sont transférées vers le serveur dans un d'objet spécial appelé groupe de modifications (*changeset*). Un groupe de modification peut contenir par exemple, la nouvelle version d'un tronçon de route qui a été déplacé, de même que les nouvelles versions d'autres tronçons autour de lui qui ont conséquemment été déplacés (Ramm et al. 2010). Il est *sérialisé* dans le format XML OSMChange. Ce format est aussi utilisé pour la diffusion régulière des mises à jour des données OSM pour ceux qui gardent une copie des données OSM dans une base de données locale (OSM 2012d). Un groupe de modifications permet à un contributeur de

[34] http://josm.openstreetmap.de/wiki/Help/Dialog/Validator
[35] http://josm.openstreetmap.de/wiki/Help/Dialog/Conflict

décrire ses changements d'une façon globale, de même il permet de tracer l'activité d'édition par contributeur, quels ont été ses changements et quand ils ont été effectués (Ramm et al. 2010). Un groupe de modifications est décrit par l'identifiant du contributeur faisant l'édition, la date d'ouverture et fermeture de l'objet (très différent de la durée de la session d'édition), le rectangle englobant les changements et un tag commentaire dont la valeur est écrite en texte libre. Un groupe de modifications contient d'abord les objets créés où chacun correspond à la première version de l'objet, puis les objets modifiés où chacun correspondant à une version successive de l'objet (version > 1), et enfin les objets supprimés où chacun correspond à la version de l'objet à supprimer. Un extrait en format XML OSMChange du groupe de modification identifié comme `10498790`[36] est présenté ci-dessous :

```
<osmChange version="0.6" generator="OpenStreetMap server">
<create>
      <node id="1605821663" lat="48.845478" lon="2.4232891"
      changeset="10498790" user="Pieren" uid="17286" visible="true"
      timestamp="2012-01-25T23:35:44Z" version="1">
            <tag k="addr:street" v="Allée des Platanes"/>
            <tag k="source" v="cadastre-dgi-fr source 2012"/>
            <tag k="addr:housenumber" v="1"/>
      </node>
</create>
<modify>
      <way id="145052755" visible="true" timestamp="2012-01-
      25T23:36:05Z" user="Pieren" uid="17286" version="2"
      changeset="10498790">
            <nd ref="1585169266"/>
            <nd ref="1585169261"/>
            <nd ref="1585169262"/>
            <nd ref="1585169267"/>
            ....
            <tag k="building" v="yes"/>
            <tag k="source" v="cadastre-dgi-fr source 2012"/>
      </way>
</modify>
<delete>
      <node id="1261703967" changeset="10498790" user="Pieren"
      uid="17286" visible="false" timestamp="2012-01-25T23:36:13Z"
      version="2"/>
</delete>
</osmChange>
```

Le nœud (version 1) et le chemin (version 2) ci-dessus identifiés comme `1605821663` et `145052755` peuvent être visualisés sur la Figure 16 (à g. et à d. respectivement).

[36] http://www.openstreetmap.org/api/0.6/changeset/10498790/download

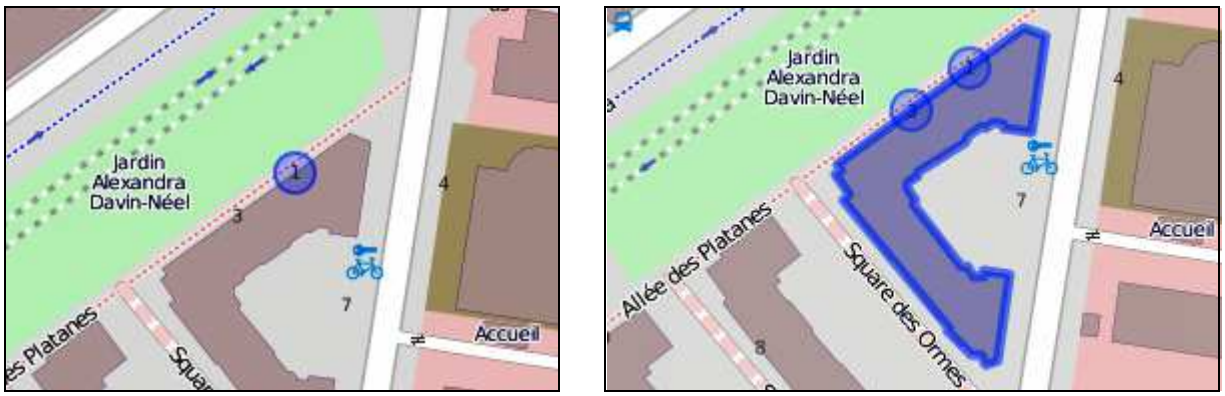



**Figure 16 : la version 1 du nœud `1605821663` et la version 2 du chemin `145052755` sur OSM**

L'élément le plus important d'un groupe de modifications est le numéro de la version de chaque objet, nécessaire pour le serveur afin d'identifier les conflits de versions (OSM 2012a). C'est-à-dire, un même objet peut être modifié par plusieurs groupes concurrents de modifications, si les versions sont différentes (*optimist locking*). Néanmoins, la même version d'un objet ne peut pas être modifiée dans plusieurs groupes différents de modifications. Si le numéro de la version fourni par le client n'est pas le même que celui du serveur, une erreur est retournée au client : `HTTP status code 409: Conflict`. Un autre conflit de versions et d'intégrité détecté par le serveur peut être par exemple celui de la suppression d'un nœud utilisé par un chemin ou d'un nœud d'un chemin (et non le chemin lui-même). Dans ce cas, le serveur retourne au client : `HTTP status code 412 Precondition failed`. Si aucun conflit n'est produit, le serveur stocke chaque nouvelle version des objets concernés dans la base de données OSM. Pour appliquer un groupe de modifications dans une transaction de base de données, il faut utiliser la méthode `diff upload` de l'API OSM (OSM 2012a).

## 3.4 La base de données OSM

Le serveur OSM accède à la base de données centrale PostgreSQL/PostGIS contenant les données OSM. Elle contient une table pour les versions actuelles des objets par leurs types avec leurs tags correspondants. En particulier, les nœuds OSM et leurs tags sont stockés, respectivement dans `current_node` et `current_node_tags`. Les chemins OSM, leurs tags et leurs nœuds sont conservés respectivement, dans `current_ways`, `current_way_tags` et, `current_way_nodes`. Les relations OSM, leurs tags et leurs membres sont stockées respectivement, dans `current_relations`, `current_relation_tags` et `current_relation_members`. Pour chacune de ces tables, il existe également des tables historiques avec des champs équivalents plus un champ version, c'est-à-dire `node_tags`, `nodes`, `way_nodes`, `way_tags`, `ways`, `relation_members`, `relation_tags`, `relations`. La base de données OSM contient aussi deux tables `changesets` et `changesets_tags` pour stocker les informations sur les groupes de modifications et leurs tags. Aucun client ne peut accéder directement à la base de données via des requêtes SQL mais uniquement via l'API OSM. Un contributeur peut toujours naviguer sur la carte OSM, cliquer sur un objet et visualiser son historique. Il est possible d'obtenir toutes les versions de l'objet et ses tags par l'URL `http://www.openstreetmap.org/browse/<typed'objet>/<id>/history`. L'exemple

ci-dessous montre en format OSM XML, deux versions[37] (#1 et #2) de l'objet chemin identifié comme `145052693` représentant le bâtiment de l'IGN à Saint-Mandé (celui en figure 9).

```
<osm version="0.6" generator="OpenStreetMap server">
      <way id="145052693" timestamp="2012-01-10T22:50:15Z" user="Pieren"
version="1" changeset="10356427">
            <nd ref="1503865153"/>
            …
            <tag k="construction" v="yes"/>
            <tag k="landuse" v="construction"/>
            <tag k="source" v="knowledge"/>
      </way>
      <way id="145052693" timestamp="2012-01-25T23:36:10Z" user="Pieren"
version="2" changeset="10498790">
            <nd ref="1503865153"/>
            …
            <tag k="construction" v="yes"/>
            <tag k="landuse" v="construction"/>
            <tag k="source" v="knowledge"/>
      </way>
</osm>
```

## 3.4 La gestion de la qualité dans OSM

La communauté OSM participe activement à la gestion de la qualité des données OSM. En effet, 58,6% des contributeurs sondés sont impliqués dans la correction d'erreurs dans les données. Certaines erreurs sont signalées par des contributeurs sur OpenStreetBugs[38], et d'autres sont détectées automatiquement par des outils Web comme Keepright[39] et Osmose[40], mis en place par la communauté de développeurs OSM (OSM 2012e). Ces outils offrent une carte en ligne afin de permettre la visualisation de telles erreurs. Un contributeur peut signaler sur ces outils qu'une erreur est déjà corrigée (par lui ou par un autre contributeur) ou qu'une erreur a été mal détectée (pour un outil ou par un autre contributeur), comme un faux positif. OpenStreetBugs (OSB) permet de signaler une erreur en texte libre en cliquant sur une localisation spécifique sur un fond de carte OSM (voir la Figure 17 à g.). OSB permet à un contributeur surveillant régulièrement une zone géographique de s'abonner à un flux RSS sur cette zone-là afin d'être automatiquement notifié des nouvelles erreurs signalées et résolues. Keepright et Osmose exécutent périodiquement, sur un extrait d'une copie locale de la base de données OSM, des procédures prédéfinies pour la détection de certaines incohérences. Par exemple, la méthode de détection d'incohérences concernant « les routes ne chevauchent pas les bâtiments » vérifie périodiquement la présence de superpositions entre des objets avec des géométries polygonales, l'un des objets possède est identifié comme un bâtiment grâce au tag `building` et l'autre objet est identifié comme une route grâce au tag `highway`. La Figure 17 (à d.) montre une incohérence concernant l'intersection de deux tronçons de route sans un nœud de jonction sur Keepright. D'autres types d'incohérences sont : une aire non fermée, des nœuds trop proches entre eux mais ne faisant pas partie du même chemin, etc.

[37] http://www.openstreetmap.org/api/0.6/way/145052693/history

[38] http://openstreetbugs.schokokeks.org/

[39] http://keepright.ipax.at/

[40] http://osmose.openstreetmap.fr/map/

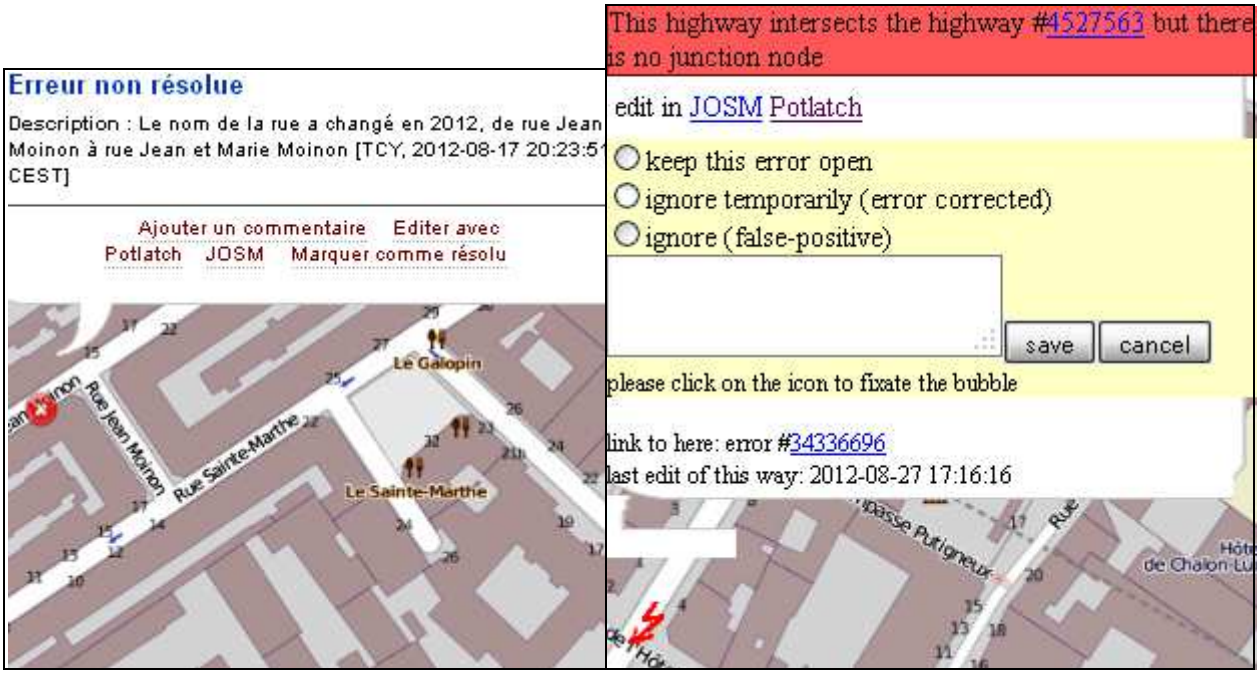



**Figure 17 : (g.) Une erreur signalée (symbole rouge) par un contributeur sur OSB et, (d.) une erreur détectée automatiquement par Keepright (symbole rouge)**

La correction d'erreurs est une tâche des contributeurs. Une fois l'erreur signalée par un contributeur ou par un des outils, un contributeur utilise un éditeur OSM (ex : JOSM) afin de le régler. Par exemple, la Figure 18 montre la séquence d'actions pour la correction, en utilisant l'éditeur en ligne Potlatch, de l'erreur identifiée par Keepright comme 34336696[41] signalant l'absence d'un nœud de jonction entre l'intersection de deux tronçons de route. Plus précisément, la Figure 18 (à g.) expose l'absence d'un nœud qui est mal placé car il n'est pas sur la « vraie » intersection mais proche d'elle. La Figure 18 (au m.) montre l'ajout du nouveau nœud de jonction dans l'intersection « vraie » intersection des deux tronçons. La Figure 18 (à d.) présente la suppression du nœud initial qui était mal placé.

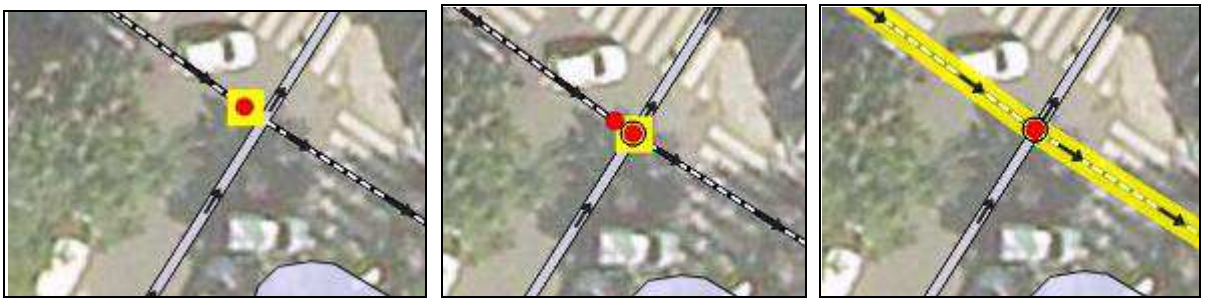


**Figure 18 : (g.) Absence d'un nœud d'intersection entre les tronçons, (m.) ajout du nouveau nœud et (d.) suppression du nœud initial mal placé**

Certaines erreurs peuvent être facilement résolues comme celui de la Figure 18. Néanmoins, d'autres erreurs seront difficiles à résoudre, comme celle en Figure 19 (à g.) détectée par

[41] http://keepright.ipax.at/report_map.php?schema=77&error=34336696

Osmose (un outil similaire à Keepright) signalant une très petite intersection entre deux bâtiments (surfaciques). La résolution peut être donc une opération délicate même pour un contributeur avec une bonne maîtrise des outils d'édition. La Figure 19 (à d.) met en évidence la difficulté de déplacer dans la bonne mesure avec un éditeur OSM (ici, Potlatch) un des deux bâtiments afin d'enlever l'intersection.

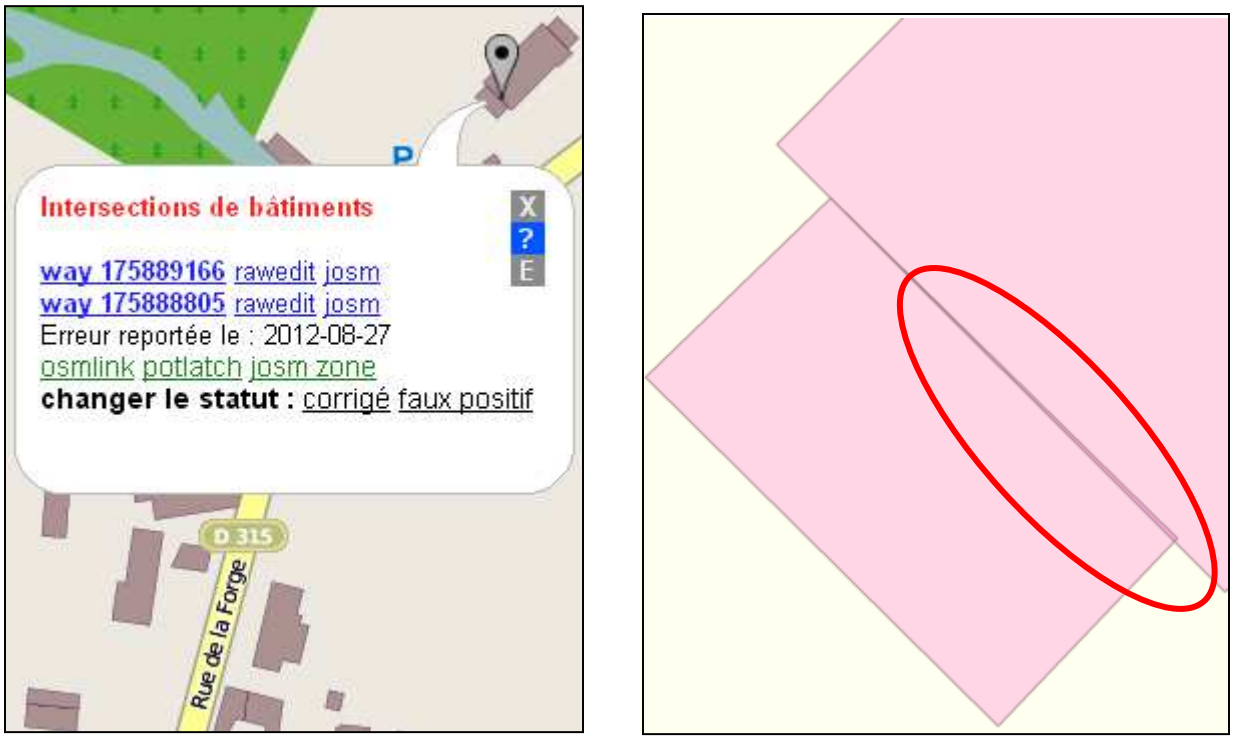



**Figure 19 : (g.) Erreur d'intersection entre deux bâtiments détectée par Osmose, (d.) l'intersection (très réduite) entre les deux bâtiments visualisée sur l'éditeur OSM Potlatch**

De plus, Osmose permet la recherche d'incohérences par région en France et par type sur l'interface Web. La Figure 20 montre les résultats de la recherche d'incohérences par type (intersections de bâtiment, bâtiments trop petits, répétition de nœuds, grosses intersections de bâtiments) sur la région de la Basse-Normandie.

france_basse_normandie | Choisir

| # | source | cl | | # | Item | titre | nombre |
|---|---|---|---|---|---|---|---|
| 96 | osmosis_building_overlaps-france_basse_normandie | 1 | | 0 | bâtiments se recouvrant | Intersections de bâtiments | 70 |
| 96 | osmosis_building_overlaps-france_basse_normandie | 2 | | 0 | bâtiments se recouvrant | Grosses intersections de bâtiments | 72 |
| 96 | osmosis_building_overlaps-france_basse_normandie | 3 | | 0 | bâtiments se recouvrant | Bâtiments trop petit | 1 |
| 96 | osmosis_building_overlaps-france_basse_normandie | 4 | | 0 | bâtiments se recouvrant | Interstice entre les bâtiments | 63 |
| 96 | osmosis_building_overlaps-france_basse_normandie | 5 | | 0 | bâtiments se recouvrant | Groupe de Grosses intersections de bâtiments | 0 |
| 89 | sax-france_basse_normandie | 103 | | 1010 | nœud répété | Répétition de nœuds | 4 |

**Figure 20 : interface Web Osmose permettant la recherche d'incohérences par région et par type**

# 4 Discussion

Cette section présente une discussion comparant plusieurs points communs (police en vert sur le Tableau 2) et différences (police en rouge sur le Tableau 2) de la production collaborative de données géographiques du point de vue d'un producteur institutionnel comme l'IGN et du point de vue d'un projet communautaire de cartographie en ligne comme OSM.

| **Critère / Producteur** | **Institutionnel** | **Communautaire** |
|---|---|---|
| Spécifications de la base de données | - **très exhaustives** et<br>- rédigées en **texte libre**. | - **peu de spécifications** dans le sens institutionnel, mais certaines parties de la documentation en ligne sont bien détaillées, et<br>- rédigées en **texte libre** facilement en ligne sur un wiki et donc très actuelles. |
| Evolution des spécifications | - les spécifications **changent fréquemment**, et<br>- le **manque de représentation en langage formel** informatique cause donc des problèmes dans l'entretien de leur évolution. | - les spécifications **changent fréquemment**, et<br>- le **manque de représentation formelle** fait qu'il est difficile d'entretenir leurs évolutions. |
| Échanges client serveur | - les **nouvelles versions des objets sont transférées** depuis le client vers le serveur. | - les **nouvelles versions des objets sont transférées** depuis le client vers le serveur. |
| Gestion de l'historique des objets de la base de données | - chaque **version d'un objet est conservée** dans la BDUni, et<br>- pour savoir ce qui a changé entre deux versions d'un objet, il faut recourir à **des opérations de comparaison** des valeurs d'attributs entre deux versions de l'objet. | - chaque **version d'un objet est gardée** dans la base de données centrale OSM, et<br>- pour savoir ce qui a changé entre deux versions d'un objet, il faut recourir à **des opérations de comparaison** des valeurs d'attributs entre deux versions de l'objet. |
| Détection et résolution d'incohérences dans les données au niveau de la géométrie ou de la topologie | - les méthodes de validation disponibles sont **très nombreuses**,<br>- les incohérences sont détectées par des **procédures de contrôle** déclenchées par l'opérateur sur son poste client,<br>- des **incohérences ne sont pas introduits** dans la BD Uni, et<br>- la correction d'incohérences est **réalisée manuellement** par l'opérateur. | - **peu de méthodes** de validation existent à l'heure actuelle,<br>- les incohérences sont identifiées automatiquement par des modules sur les clients OSM (principalement JOSM) exécutant des **méthodes de validation**,<br>- certaines **incohérences sont détectées *a posteriori* automatiquement** dans la BD OSM par les outils de gestion de qualité, et<br>- la correction d'incohérences est **réalisée manuellement** par le contributeur. |

| | | |
|---|---|---|
| Détection et résolution de conflits de versions entre les individus | - la **détection est automatiquement** gérée par le système de gestion de versions GCVS,<br>- la définition d'une zone de réconciliation aide à **contrôler l'édition et diminuer les conflits**,<br>- le stockage des informations sur la réconciliation dans la BDUni aide à identifier manuellement et comprendre les changements faits par chaque opérateur impliqués dans le conflit,<br>- la résolution est **manuelle**. | - la **détection est automatiquement** gérée par le serveur OSM,<br>- la définition d'un groupe de modifications et le stockage de leurs informations dans la BD OSM **aide à identifier manuellement les changements** faits par chaque contributeur impliqué dans le conflit, et<br>- la résolution est **manuelle**. |
| Évaluations de qualité des données et leur documentation | - sont **régulières et obligatoires** sur le territoire, et<br>- les résultats **font partie** de la documentation et des métadonnées de chaque produit IGN. | - sont effectuées par la communauté de recherche de façon **très irrégulière**, et<br>- **ne sont pas incorporées** dans la documentation OSM ou comme métadonnée. |
| Licence des données | - **propriétaire** mais gratuite pour la recherche. | - licence **libre** CC-By-Sa. |

**Tableau 2 : tableau comparatif montrant les ressemblances et les divergences entre les modes de production collaborative de données géographiques institutionnels et communautaire**

Globalement, nous distinguons plus de ressemblances que de divergences. Parmi les ressemblances remarquables, la spécification des bases de données IGN et certaines parties importantes de la documentation comme la page des Map Features sont produites et diffusées en texte libre. Dans les deux cas, cela est une contrainte pour gérer l'évolution de cette documentation dans le temps car elle change très fréquemment. Il est donc difficile de les interroger et de les entretenir de façon simple et automatisée. Par exemple, dans le cas OSM, il serait intéressant d'avoir une réponse à la question suivante : quels sont les tags concernés par un thème sur les zones humides et comment certains tags ont-t-ils changé ? Une divergence est l'exhaustivité des spécifications de l'IGN. En revanche, uniquement certaines parties de la documentation OSM pouvant être considérées comme une « spécification », sont très bien documentées (ex : la page des *Map Features*) et mises à jour.

Une autre ressemblance est la structure du message échangé par le client vers le serveur, c'est-à-dire la nouvelle version de l'objet (et non seulement ce qui a changé) est transférée vers le serveur. De la même manière, la façon dont l'historique de données est stocké est ressemblante. Il n'est donc pas possible de savoir facilement ce qui a changé entre deux versions d'un objet, il faut recourir à des opérations de comparaison des valeurs d'attributs entre deux versions de l'objet. Un outil comme OSM History Browser[42] vise à servir de support visuel pour le contributeur pour consulter les différences entre deux versions d'un objet.

[42] http://osm.virtuelle-loipe.de/history/

Une autre ressemblance est la détection des incohérences. Dans les deux cas, des incohérences sont automatiquement détectées par des méthodes de validation exécutées par le client sur une copie locale du contenu. Les méthodes de validation disponibles sont très nombreuses dans le cas de l'IGN : une équipe de développeurs est exclusivement dédié à cette tache. Le but est de ne pas introduire d'incohérences dans la BDUni. Dans le cas OSM, il n'y a pas de contrôles sur ce qui est mis dans la base données. Il est donc nécessaire de mettre en place des outils pour la détection d'incohérences, *a posteriori*, dans la BD OSM (ex : Keepright). Dans les deux cas, la correction d'incohérences est manuelle. De la même façon, les conflits de versions sont automatiquement détectés dans les deux cas, de la même manière qu'un gestionnaire de versions comme SVN ou Git le fait. Néanmoins, certains "conflits" ne devraient pas être détectés comme tels, car la modification d'une géométrie et d'un attribut thématique d'un objet par deux collecteurs ne devrait pas être considérée comme un vrai conflit. C'est-à-dire, un conflit n'est qu'intrinsèquement lié à l'édition d'une même version par deux individus différents. Du côté IGN, la définition d'une zone de réconciliation aide à contrôler l'édition et à diminuer les conflits. Dans les deux cas, le stockage des informations sur la réconciliation dans la BDUni et sur les groupes de modifications dans la BD OSM aide à identifier manuellement et comprendre les changements faits par chaque contributeur impliqué dans le conflit. Dans les deux cas, la résolution est manuelle et peut être guidée par un outil visuel pour aider à la décision.

Une différence très importante est l'assiduité des évaluations de qualité et de leur documentation. A l'IGN, les évaluations de qualité sur toutes les bases de données sont régulières, obligatoires, exhaustives, et leur documentation est incorporée comme métadonnée dans le produit IGN suivant la norme ISO 19115 (ISO 2003). De l'autre coté, les chercheurs en sciences de l'information géographique s'intéressent à évaluer la qualité des données OSM en utilisant une source de référence, Haklay (2010) au Royaume-Uni, et Girres et Touya (2010) en France. Néanmoins, certaines de ces évaluations n'ont pas encore été intégrées dans OSM comme par exemple l'ajout d'une couche de visualisation de la qualité des données. Une dernière différence importante est l'incompatibilité des licences, un sujet complexe, d'une grande importance et très actuel puisqu'il empêche la prolifération d'échanges entre les deux modèles de production.

Nous soulignons pour conclure un besoin dans ces modes de production qui nous semble important et sur lequel nous focaliserons notre proposition : celui de la *gestion de la cohérence d'un contenu géographique construit de manière collaborative*. Il y a un intérêt d'assurer la cohérence des données afin de permettre que des prises de décision s'appuient dessus. Nous nous posons donc la question de « *que veut dire la cohérence* » d'un contenu géographique collaboratif. Nous cherchons à répondre à cette question dans le chapitre suivant sur l'état de l'art (chapitre II).

# Chapitre II

# État de l'art : l'édition collaborative, la gestion de la qualité et de la cohérence

Ce chapitre recense des travaux de recherche liés à la problématique de ce travail de thèse : l'édition collaborative d'un contenu et la gestion de sa qualité, en particulier la gestion de sa cohérence.

En section 1, nous investiguons les mécanismes et les structures d'organisation de l'information dans un éditeur collaboratif, par exemple un moteur de wiki et un éditeur collaboratif de documents XML, utilisés pour gérer semi-automatiquement la cohérence du contenu.

En section 2, nous nous intéressons à des recherches sur des données géographiques issues de projets communautaires intéressés à construire des contenus libres, aussi dits VGI (pour rappel, acronyme de *Volunteered Geographic Information*). Nous étudions, les aspects concernant le problème de la qualité des données et la gestion de la cohérence au niveau de la caractérisation des entités géographiques.

En section 3, nous décrivons des recherches visant à améliorer la cohérence des données géographiques pour la gestion et l'amélioration de leur qualité, de même des travaux sur la problématique d'appariement et transformation de données géographiques provenant de sources différentes pour leur intégration.

Enfin, en section 4, nous faisons un bilan où nous identifions dans ces travaux, les éléments pertinents pour notre proposition de thèse.

## 1 La gestion de la cohérence dans un éditeur collaboratif

Cette section décrit les éléments de gestion de la cohérence d'un contenu non-géographique dans un éditeur collaboratif. Un éditeur collaboratif facilite l'édition d'un contenu commun par un groupe de personnes qui travaillent à des moments et à des emplacements différents (Oster et al. 2007). Quelques bons exemples sont les moteurs de wiki et les éditeurs collaboratifs de modèles UML. Un éditeur collaboratif facilite la construction d'un contenu commun en encourageant l'utilisation d'un vocabulaire afin d'enlever les incohérences de sens pendant l'édition comme par exemple un article d'un wiki, ainsi que pour l'organisation de l'information décrite dans cet article. De plus, un éditeur collaboratif décompose le contenu en considérant la notion de structure. Afin de gérer les incohérences, cette décomposition permet de rendre les éditions sur des parties différentes du contenu indépendantes.

## 1.1 Utilisation d'un vocabulaire pour éviter les incohérences de sens et organiser l'information

Un vocabulaire permet de définir sans ambiguïté le sens d'un mot important. Il est possible d'utiliser le même mot tout au long du contenu collaboratif afin de désigner un concept. Un vocabulaire permet aussi d'organiser l'information afin d'assurer la cohérence globale du contenu et faciliter son traitement automatique. Les éléments du vocabulaire à définir dépendent de la nature du contenu. Un contenu encyclopédique comme celui de Wikipédia qui vise à décrire les connaissances du monde, possède un vocabulaire composé d'instances, de leurs relations, de leurs propriétés et de leurs catégories. Ces éléments sont encodés dans le moteur de wiki sous la forme de wiki-liens, de wiki-catégories, et de modèles infobox. La Figure 21 montre ces éléments dans l'article Wikipédia sur la Place de la République à Paris. La gestion de la cohérence dans Wikipédia est une tâche des contributeurs et certaines fois, des robots (*bots*). Un *bot* est un mécanisme automatique qui effectue une vérification et des corrections si possible. A l'heure actuelle, les bots existants sont très basiques, par exemple, la plupart sont des correcteurs de fautes orthographe. Les incohérences les plus importantes sont toujours détectées et découvertes par les contributeurs de Wikipédia grâce à des éléments de vocabulaire Wikipédia.


**Figure 21 : l'entité géographique Place de la République, son lien d'homonymie (en rouge), ses propriétés et liens vers d'autres entités (en orange) et ses catégories (en vert)**

Le vocabulaire Wikipédia fournit des liens explicites, et certaines fois typés, entre deux instances. Ces relations constituent le graphe d'articles de Wikipédia (WAG) (Zesch et Gurevych 2007), où chaque nœud est un article de l'instance correspondante et chaque arc représente la relation explicite. Il existe plusieurs types de relation dans le WAG. Un lien d'homonymie clarifie le sens d'un terme (Mihalcea 2007). Par exemple, le terme `Ligne 1` possède plusieurs connotations, il existe donc une page d'homonymie intitulée `Ligne 1 (homonymie)` qui liste les nombreuses lignes dans le monde numérotées 1. Un lien de redirection permet de lister des termes alternatifs pour un même terme. Par exemple, `USA` et `Etats-Unis` possèdent la même signification, il existe donc un lien de redirection entre ces

deux articles. Par exemple, il est possible pour un contributeur de vérifier ces liens grâce à la syntaxe d'homonymie. Les autres types des liens ne sont pas explicités par les contributeurs de Wikipédia et ne sont donc pas connus. Afin d'expliciter le type d'une relation dans un wiki, des moteurs de wiki sémantiques (Krötzsch et al. 2007) proposent à des contributeurs d'ajouter une annotation dans le texte libre de l'article. Par exemple, le texte suivant (en syntaxe wiki) dans l'article sur la ville de Berlin : « `Berlin est la capitale de [[l'Allemagne|est-ville-capitale-de::Allemagne]]` », indique un lien typé `est capitale de` entre `Berlin` et `Allemagne`. Le fait d'avoir des relations typées aide à mettre en place des mécanismes automatiques de validation dans le moteur de wiki afin d'enlever les incohérences. Un contributeur peut interroger un wiki sémantique afin de vérifier, par exemple, que les capitales saisies sont correctes par rapport à la réalité et, corriger les erreurs si nécessaire.

Le vocabulaire Wikipédia permet d'établir une classification commune afin de catégoriser les entités représentées dans l'encyclopédie selon leurs natures. Dans Wikipédia, une *wiki-catégorie* désigne un concept. L'ensemble de catégories est organisé dans une structure taxonomique connue comme le graphe de catégories de Wikipédia (WCG), où chaque catégorie peut avoir une quantité arbitraire de sous-catégories établie grâce à une relation non-explicites d'hyperonymie ou de méronymie (Zesch et Gurevych 2007). Une relation d'hyperonymie est une relation entre deux termes distinguant le terme le plus général de celui plus spécifique. Une relation de méronymie distingue la partie du tout. Dans Wikipédia, la catégorie `zone humide` du WCG possède la sous-catégorie `Marais`, et la catégorie `Canal` possède la sous-catégorie `Pont-Canal`. De plus, il est possible de trouver dans le WCG des concepts « localisés » regroupant des entités d'une catégorie et localisées sur un étendu administratif (ville, région, pays), par exemple, les catégories `Fontaines du 2e arrondissement de Paris` ou `Monuments historiques de Paris`. Un des avantages de l'utilisation de catégories dans Wikipédia est de permettre à un contributeur d'accéder au contenu via le vocabulaire et de découvrir des incohérences. Prenons l'exemple d'une fontaine du 2ème arrondissement de Paris qui a été mise incorrectement dans la catégorie Wikipédia `Fontaines du 3e arrondissement de Paris`, un contributeur peut signaler et corriger l'erreur en modifiant le wiki-lien de la catégorie sur l'article.

Le vocabulaire de Wikipédia fournit des *modèles infobox* afin de décrire de façon homogène les objets appartenant à une catégorie. Certains modèles infobox correspondent à des catégories du WCG. Par exemple, de nombreux articles concernant les montagnes du monde utilisent l'infobox `Montagne` et il existe une catégorie du même nom `Montagne`. Néanmoins, la taxonomie extraite des infoboxes contient peu de concepts par rapport au WCG (Wu et Weld 2008; Nastase et al. 2010). Un modèle infobox définit des types de propriété (et ses valeurs possibles) qui peuvent être instanciés sur une entité de la catégorie correspondante. Par exemple, l'article `Place de la République` utilise le `Modèle Infobox : Voie parisienne` et définit la propriété longueur avec une valeur de 283 mètres et très récemment, des relations typées vers d'autres entités comme par exemple, la relation « voies desservies » établie entre la Place de la République et la rue du Temple. Le fait d'utiliser des modèles infobox déjà définis sur Wikipédia permet d'homogénéiser la façon dont les instances sont décrites.

## 1.2 La réconciliation fondée sur un modèle d'éditions pour réduire les incohérences

Un éditeur collaboratif décompose le contenu en considérant la notion de structure. Afin de gérer les incohérences, cette décomposition permet de rendre indépendantes les éditions sur des parties différentes du contenu. De cette façon, un mécanisme automatique peut réduire le nombre de faux conflits pendant l'édition concurrente de ces sections indépendantes par des contributeurs différents. Il peut également identifier les vrais conflits spécifiques à un type de contenu particulier (ex : un modèle UML) grâce à des contraintes qui ont été définies sur la structure du contenu. Ce mécanisme est connu dans un éditeur collaboratif sous le nom de réconciliation automatique.

Le contenu peut parfois être peu-structuré, comme par exemple le texte d'une page wiki en syntaxe wikicode. Une page wiki est décomposée en sections, paragraphes, lignes, mots et caractères. Le contenu peut être aussi (semi-)structurée, par exemple un document XML ou un modèle de classe UML. Dans ce cas, un document XML est décomposé en nœuds XML et leurs attributs. La décomposition du contenu en considérant sa structure permet de préciser sous la forme d'une édition, comment et quelle partie du contenu est changée (ex : l'ajout d'une nouvelle ligne au début de la page wiki, ou la suppression du troisième nœud du document XML). Cette décomposition permet la définition d'un modèle d'éditions qui considère les types d'éditions possibles à effectuer sur le contenu. Par exemple, Martin (2011) définit un modèle d'éditions décrivant les types d'éditions possible à appliquer sur un document XML : `ajouterNoeud(nœud, cheminArbreXML)`, `ajouterProprieteNoeud (prop, val, nœud)`, `suppressionNoeud (noeud)`, etc. Un modèle d'édition facilite la réconciliation entre deux séquences d'éditions, chacune étant effectuée séparément sur une copie différente du même contenu. La réconciliation est optimiste car toute édition sur le contenu est permise. Son objectif est d'arriver à un consensus en garantissant à terme une convergence des copies divergentes. En revanche, les approches pessimistes des systèmes de gestion de bases de données bloquent l'accès concurrent au contenu en écriture lors d'une édition effectuée par un utilisateur.

La réconciliation identifie les éditions indépendantes et conflictuelles dans les deux séquences afin de gérer les incohérences. Une paire d'éditions est indépendante quand chacune des éditions est effectuée sur une partie différente du contenu commun (Oster et al. 2006). L'édition de parties différentes du contenu, par exemple la modification de deux attributs différents d'un nœud du document XML, ne produit pas un conflit. Il est également possible d'identifier les éditions conflictuelles. Une paire d'éditions est conflictuelle quand chacune des éditions est effectuée sur la même partie du contenu commun. Par exemple, la modification du même attribut d'un nœud du document XML par deux personnes produit un conflit. Pour résoudre un conflit, elle peut offrir selon le cas, une correction automatique comme changer l'ordre des éditions ou simplifier des éditions redondantes (Michaux et al. 2011). Elle peut aussi se servir d'un historique d'éditions où toutes les opérations effectuées sont stockées afin de pouvoir revenir à un état précédent (Preguiça et al. 2003). Dans le cas contraire, elle peut notifier la

présence ou guider la résolution du conflit en levant une alerte ou en suggérant une correction afin d'être appliquée par l'utilisateur. Il est également possible de détecter une édition conflictuelle en considérant des contraintes définies au niveau du schéma, par exemple le DTD d'un document XML (Oster et al. 2007). Par exemple, l'ajout de deux nœuds `<titre>` dans le nœud `<ouvrage>` est détecté comme une incohérence si une telle contrainte est définie dans le DTD du document. Pour résoudre ce conflit, une stratégie est de choisir arbitrairement un nœud `<title>` et de mettre l'autre comme commentaire (entre `< !-- -->`) dans le document.

Plus précisément, il existe deux stratégies de réconciliation automatique de deux séquences d'éditions dans les éditeurs collaboratifs : la transformation opérationnelle et la réconciliation dite « sémantique ».

La transformation opérationnelle Ellis et Gibbs (1989) est une stratégie de réconciliation automatique se servant d'un modèle d'éditions. Elle a été mise en place dans un moteur de wiki pair à pair (P2P) (Oster et al. 2006) et un éditeur P2P de documents XML (Oster et al. 2007; S. Martin 2011). Dans un contexte P2P, il n'y a pas un contenu central. Au contraire, chaque utilisateur garde une copie du contenu. Au moment de l'édition d'une des copies, les autres doivent aussi refléter ces changements. Considérons l'exemple de Molli et al. (2003) : deux utilisateurs `user1` et `user2` qui travaillent chacun sur une copie d'une page wiki : `site1` et `site2` (voir la Figure 22). L'insertion d'un caractère `C` dans la position `P` du texte se traduit sous la forme d'édition `Ins(c,p)`. Considérons l'édition du mot `efect` qui est trouvé dans le texte et qui possède une mauvaise orthographe. L'utilisateur `user1` applique l'édition `op1=Ins(2,f)` sur la copie `site1` afin de corriger la faute. L'utilisateur `user2` applique l'édition `op2=Ins(5,s)` sur la copie `site2` afin de corriger aussi la faute. Ensuite, l'édition faite par `user1` doit être appliquée sur le contenu de `user2`. De la même manière, l'édition faite par `user2` doit être appliquée sur le contenu de `user1`. L'objectif est que les intentions des deux utilisateurs soient préservées dans chaque copie. Nous avons vu dans cet exemple que l'intention des deux utilisateurs était de corriger le mot `efect` à `effects`.

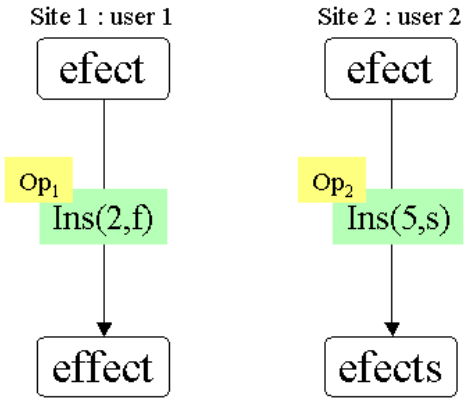



**Figure 22 : deux utilisateurs user1 et user2 qui éditent concurremment le mot efect, chacun dans sa copie locale de contenu du wiki (Molli et al. 2003)**

Ensuite, quand `op1` est exécutée sur la copie site2, nous obtenons le résultat désiré : `effects`. Par contre, quand `op2` est exécutée sur la copie `site1`, nous obtenons un résultat non-désiré :

`effecs`. Il y a donc une divergence entre les deux copies. La Figure 23 (à g.) illustre ce cas d'intégration incorrecte des éditions sur les copies. Pour résoudre ces conflits d'édition, la transformée d'opérations est un concept clé dans la réconciliation. Une transformée d'opération est une fonction qui prend deux éditions affectant la même version du contenu et retourne une nouvelle édition. Un exemple d'une transformée qui sait gérer les conflits d'éditions est défini ci-dessous (en pseudocode) :

```
T( Ins(p1, c1), Ins(p2, c2) ) :-
if ( p1 < p2 ) then
      return Ins(p1, c1)
else
      return Ins(p1+1, c1)
endif
```

La Figure 23 (à d.) illustre la façon dont cette transformée peut résoudre le conflit d'édition de l'exemple précédent. Pour la copie `site1`, la transformée `T(Ins(5,s),Ins(2,f))` renvoie la valeur `Ins(6,s)`. Le résultat obtenu est bien celui désiré : `effects`. Pour la copie `site2`, la transformée `T(Ins(2,f),Ins(5,s))` renvoie `Ins(2,f)`, ce qui aboutit au résultat souhaité : `effects`.

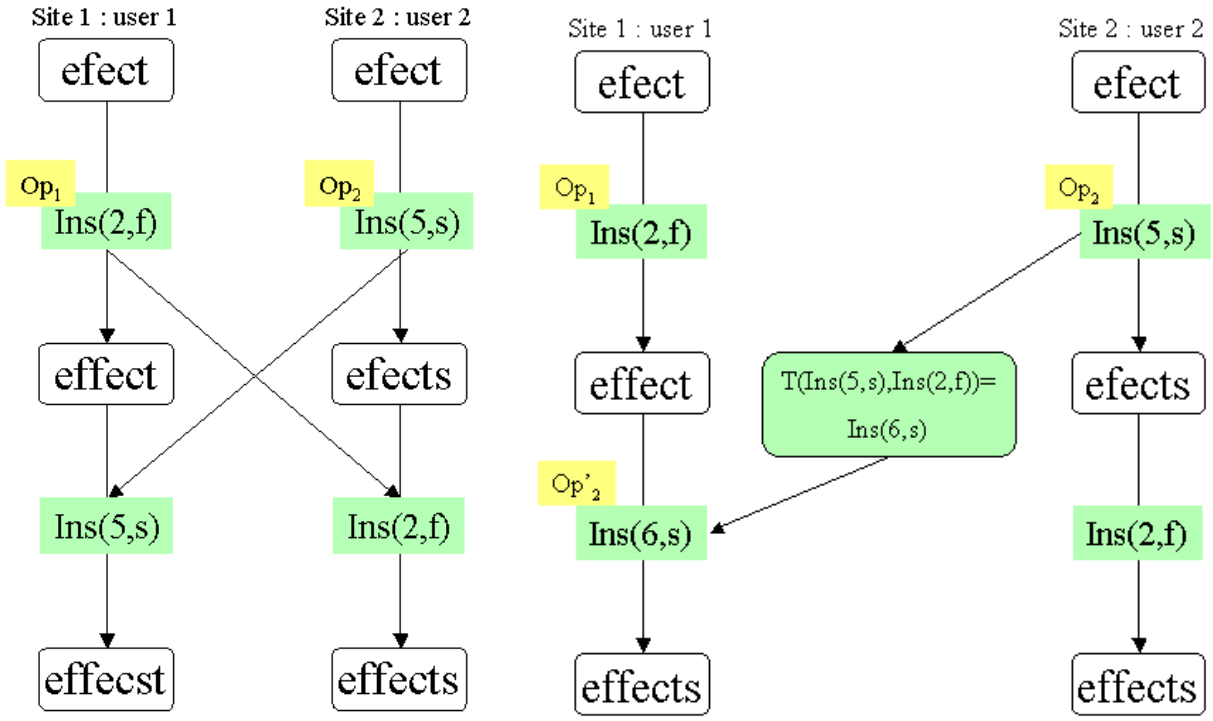



**Figure 23 : exemple d'une transformée d'opérations (Molli et al. 2003), (à g.) l'intégration incorrecte de deux séquences d'éditions, et (à d.) la réconciliation correcte de ces deux séquences**

La réconciliation dite « sémantique » est une autre stratégie de réconciliation se servant aussi d'un modèle d'éditions. Cette approche a été conçue et utilisée pour un agenda partagée en ligne (Edwards et al. 1997; Preguiça et al. 2003) dans un environnement centralisé, et pour un éditeur collaboratif de modèles UML (Michaux et al. 2011). La réconciliation sémantique repose

sur la définition de règles qui doivent être respectées afin que le contenu reste cohérent. Certains règles sont génériques et d'autres sont spécifiques au contexte applicatif. Selon (Michaux et al. 2011), il existe trois règles génériques qui doivent toujours être vérifiées. La première règle est la causalité qui est une relation d'ordre partiel entre les éditions (Lamport 1978). Prenons par exemple un élément modifié dans une séquence d'opérations, la création de cet élément devra s'effectuer avant la modification. La deuxième règle est la suppression cohérente des éléments, c'est-à-dire que dans le cas où un élément est supprimé dans une séquence d'éditions, il n'est pas possible que cet élément soit modifié après la suppression. La dernière règle est l'existence des identifiants uniques dans tout le contenu qui est indispensable pour faire référence à chaque élément de manière non ambigüe. Il est possible de définir des contraintes spécifiques à une application afin de définir les conflits potentiels d'édition. Tout d'abord, il est nécessaire de définir les opérations élémentaires qui peuvent être effectuées sur le modèle. Pour un modèle UML, Michaux et al. (2011) définissent les opérations suivantes :

`create (me,mc)` crée un élément du modèle `me`, instance de la méta-classe `mc`,
`delete (me)` supprime un élément du modèle `me`,
`addProperty (me,p,v)` assigne la valeur `v` à la propriété `p` de l'élément du modèle `me`,
`remProperty (me,p)` supprime la valeur, le cas échéant, de la propriété `p` de l'élément du modèle `me`,
`addReference (me,r,met)` assigne à l'élément du modèle `met`, la référence `r` avec l'élément du modèle `me`,
`remReference(me,r)` supprime la valeur, le cas échéant, de la référence `r` de l'élément du modèle `me`.

Considérons l'exemple de  concernant deux utilisateurs `user1` et `user2` qui éditent un modèle UML de manière collaborative. `User1` applique les éditions suivantes :

```
ed1. create(c1,Class),
ed2. addProperty(c1,name,'Apple'),
ed3. create(c2,Class),
ed4. addProperty(c2,name,'Fruit'),
ed5. create(a1,Attribute),
ed6. addProperty(a1,name,'variety'),
ed7. addProperty(a1,type, 'String'),
ed8. addReference(c1,attribute,a1),
ed9. addReference(c1,super,c2).
```

La Figure 24 montre le modèle UML correspondant aux éditions effectuées par `user1`.

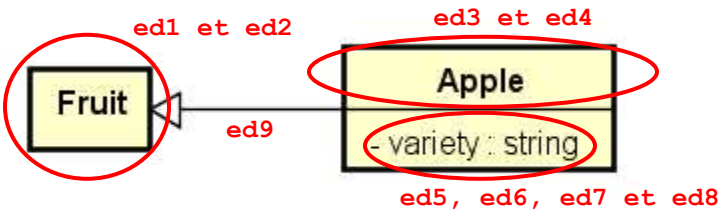



**Figure 24 : modèle UML correspondant aux éditions effectuées par `user1`**

`User2` reçoit les éditions d'`user1`. A ce moment, les deux utilisateurs possèdent donc des copies exactes du modèle UML. Ensuite, `user1` effectue encore quelques éditions pour ajouter une propriété à la classe `Apple`, listées ci-dessous :

```
ed10. create(a2,Attribute),
ed11. addProperty(a2,name,'nbPips'),
ed12. addProperty(a2,type, 'int'),
ed13. addReference(c1,attribute,a2).
```

La Figure 25 montre le modèle UML correspondant au deuxième groupe d'éditions effectuées par `user1`.

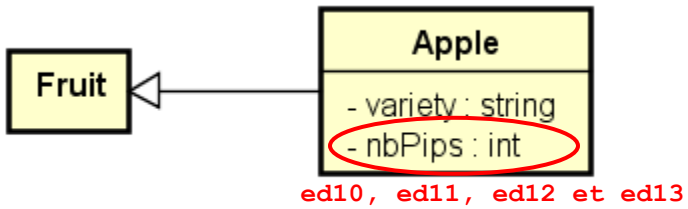



**Figure 25 : modèle UML correspondant au deuxième groupe d'éditions effectuées par `user1`**

Le système considère pour l'instant `user1` comme l'utilisateur le plus productif car il a effectué plus d'opérations sur le modèle.

De son côté, `user2` supprime la classe `Apple`. Les deux utilisateurs possèdent maintenant deux copies divergentes du modèle. Quand `user2` essaie de récupérer les éditions d'`user1`, plusieurs conflits sont détectés par le système. Plus précisément, chaque modification sur la classe `Apple` effectuée par `user1` est en conflit avec la suppression effectuée par `user2`. Pour résoudre ce conflit, la réconciliation dite sémantique donne la priorité aux éditions effectuées par l'utilisateur le plus productif : `user1`. C'est-à-dire, les modifications effectuées par `user1` seront prises en compte et non la suppression faite par `user2`. Néanmoins, il est possible plutôt de garder les suppressions faites par `user2` comme une alternative. En tout cas, les éditions des deux utilisateurs sont conservées dans l'historique afin d'explorer les alternatives et pouvoir revenir en arrière.

# 2 La gestion de la cohérence et de la qualité de contenus géographiques collaboratifs dits VGI

Cette section passe en revue des travaux en sciences de l'information géographique portant sur les contenus géographiques collaboratifs, dits VGI (Goodchild 2007). La Figure 26 présente temporellement des papiers clés (et leurs liens de référencement bibliographique) qui expliquent le phénomène VGI. Ces travaux signalent également les nouveaux défis et exposent des nouvelles directives de recherche en sciences de l'information géographique concernant le VGI : il faut concevoir des nouvelles méthodes adaptées à l'exploitation de ces données (Kuhn 2007). Cette revue se focalise sur les aspects de la gestion de la cohérence du VGI (et de leur modèle), ainsi que sur l'évaluation de la qualité de ces données, principalement pour des données Open Street Map car les recherches sont de plus en plus nombreuses et actives.

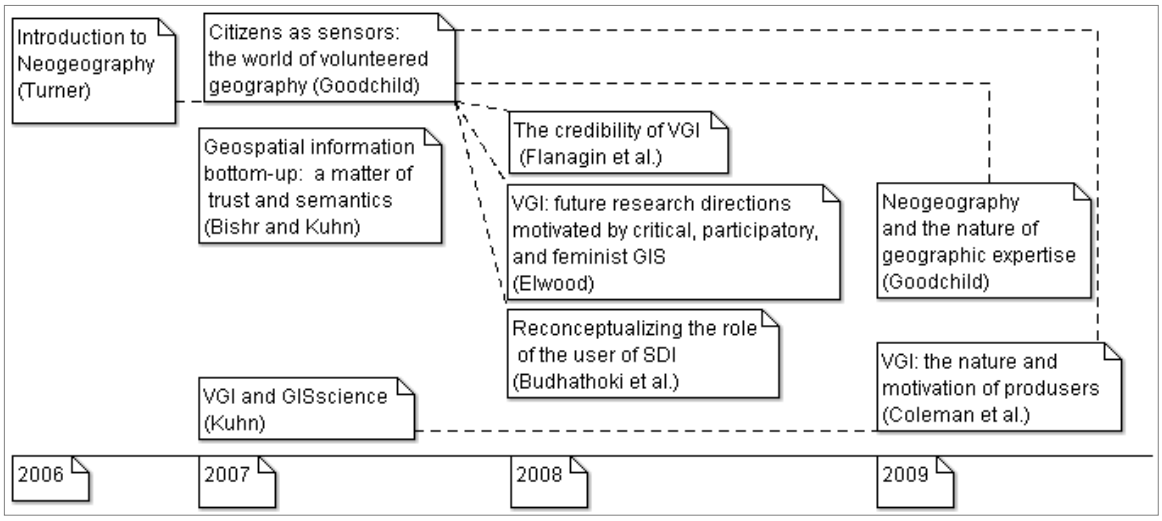



**Figure 26 : quelques papiers clés qui définissent le phénomène VGI indiquant les nouvelles directives de recherche des sciences de l'information géographique concernant le VGI**

## 2.1 La gestion de la cohérence au niveau de la caractérisation d'entités

Dans un projet communautaire de cartographie collaborative, un objet est décrit en utilisant des étiquettes (*tags*). Ces tags sont conçus librement par les contributeurs. Cette collection de mots-clés est connue autrement avec l'appellation de folksonomie (Limpens et al. 2009). Dans Open Street Map (OSM), ces tags sont de la forme `clé=valeur`, et certaines fois, sont documentés en texte libre sur la page Map Features. La charte de directives du projet encourage les contributeurs à les réutiliser. Néanmoins, les contributions n'adhèrent pas systématiquement à ce modèle de tags lors de la caractérisation des objets (Girres et Touya 2010). En effet, l'utilisation d'un schéma ou d'une taxonomie formelle n'est pas une pratique commune d'un projet communautaire (Bishr et Kuhn 2007; Exel et al. 2010). Dans OSM, l'adhérence à des tags par les contributeurs n'est pas vérifiée au niveau de l'interface d'édition des données ni au niveau de la base de données OSM (via l'API). Mooney et Corcoran (2011) signalent un manque important d'adhérence des tags par les contributeurs et la présence notable d'erreurs de frappe. Ils montrent par exemple qu'il existe 29 valeurs différentes pour le tag `landuse` proposées dans la page Map Features, et que pour 577 objets contenant ce tag, il y a 10 valeurs qui sont inconnues par la communauté. Cette liberté amène une difficulté importante pour les contributeurs : comment choisir le tag le plus approprié pour bien décrire un objet. Ce problème touche globalement à la cohérence interne des données (Exel et al. 2010). En revanche, Kuhn (2007) exprime le besoin de permettre l'émergence de nouvelles connaissances de la part de ces communautés au lieu de prédéfinir des schémas ou des taxonomies formelles pour contrôler la caractérisation de ces objets. En effet, il existe un besoin de capturer les multiples perceptions des contributeurs de VGI sur des lieux, c'est-à-dire, comment ces gens perçoivent un lieu qui possède plusieurs significations (selon le point de vue ou la perspective) et qui évolue dans le temps (Roche et Feick 2012). Ainsi, il est possible d'analyser ces connaissances afin de les transformer en structures formelles. En particulier, Purves et al. (2011) s'intéressent à explorer la folksonomie issue du projet communautaire

Flickr[43] afin de quantifier comment les contributeurs décrivent fréquemment l'espace à travers des tags. Un de leurs résultats est que la folksonomie de Flickr est une source particulièrement riche de termes pour décrire des événements et des activités humaines. Ces études peuvent aider à mettre en place des mécanismes d'aide aux contributeurs comme par exemple la réutilisation et la suggestion de tags pertinents afin de trouver un compromis entre la cohérence des données et la liberté de saisie accordée à des contributeurs. Les recherches actuelles proposent différents mécanismes et formalismes pour améliorer la gestion de la cohérence au niveau de la caractérisation d'entités pour un projet communautaire comme OSM.

Antoniou (2011) propose un mécanisme à la volée pour la validation pendant la contribution de l'adhérence des contributions à un schéma prédéfini. D'abord, l'auteur transforme semi-automatiquement les tags OSM spécifiés dans la page Map Features sous la forme d'un schéma XML et ensuite propose d'utiliser ce schéma prédéfini dans l'outil d'édition des données afin de valider à la volée, l'adhérence des contributions à ce schéma. Une interface graphique d'édition, comme celle de la Figure 27, propose un formulaire par objet correspondant à des clés de tags, où le contributeur saisit chaque valeur. Le système valide si la valeur fournie par le contributeur correspond au domaine de valeurs défini dans le schéma XML et si son format est correct. Par exemple, un contributeur qui saisit un code postal ne respectant pas le format de codes postaux (alphanumériques en Royaume-Uni), recevra une notification lui indiquant l'erreur.


**Figure 27 : prototype de la proposition d'(Antoniou 2011) pour contrôler le processus de tagging dans OSM**

Mülligann et al. (2011) proposent un mécanisme de suggestion d'éléments du schéma (ex : des types de *features*) par analyse de contexte spatial afin d'aider les contributeurs à correctement caractériser les objets. Les auteurs analysent l'historique de données OSM afin de trouver des

[43] http://www.flickr.com/

patrons spatio-temporels des points d'activité ou intérêt (PAI) dans l'espace. Cet historique permet également de concevoir une mesure de similarité sémantique entre leurs tags. Il est donc possible de déterminer pendant la saisie d'un PAI, quel est son tag le plus probable en considérant son espace géographique et les autres PAIs autour cet objet. Par exemple, un objet avec le tag `amenity=bar` est probablement près d'autres objets PAI avec un tag `amenity=nightclub`, ainsi que son horaire d'ouverture est dépendant des horaires des autres PAI autour lui. Un tel mécanisme peut aider un contributeur à l'aide d'alertes afin d'éviter d'introduire des informations incohérentes par rapport au monde réel.

Scheider et al. (2011) proposent de changer la manière dont les contributeurs assignent des tags à des points d'intérêts et d'activité (PAI). Plus précisément, un PAI peut être décrit selon ses multiples fonctions dans le monde réel (ex : un restaurant sert à boire, à manger, propose des toilettes, etc.). Donner des descriptions exhaustives sur la fonction des objets n'est pas possible dans OSM. Les auteurs proposent de formaliser ces descriptions sous la forme de clauses de Horn (Sterling et Shapiro 1994)[44] afin de faciliter la détection automatique des informations incohérentes en évaluant ces règles. De cette façon, il est possible d'exprimer la règle présentée en Équation 1 pour décrire à un instant `t`, certains PAIs (ex : bars, cafés et restaurants) comme des endroits où les gens peuvent boire et manger. Ainsi, un mécanisme automatique peut trouver des informations incohérentes par rapport à un ensemble de règles, par exemple un bar fermé le vendredi ou le samedi soir (puisque cela n'est en général pas le cas).

```
Place(poi) ∧ ∃eatingplace.P(eatingplace, poi) ∧ ∃somebody, something,
t.Affords(eatingplace, doEat(somebody, something, t)) ∧
∃drinkingplace.P(drinkingplace, poi) ∧ ∃somebody,
t.Affords(drinkingplace, doDrinkAlcohol(somebody, t))
```

**Équation 1 : expression en logique de Horn décrivant les fonctions de certains PAIs dans le monde réel : des bars, cafés et restaurants sont des endroits où les gens peuvent manger et boire (Scheider et al. 2011)**

En revanche, Codescu et al. (2011) et Ballatore et al. (2012) ne cherchent pas à changer le processus de caractérisation des objets dans OSM. Ils structurent les connaissances existantes dans les tags OSM sous la forme, respectivement, d'une ontologie et d'un graphe RDF, afin de faciliter la détection d'incohérences dans le modèle de tags (ex : unifier des concepts redondants, détecter des conceptualisations conflictuelles). Les incohérences au niveau du modèle peuvent être manuellement résolues *a posteriori* par les contributeurs et le résultat incorporé ensuite dans OSM. Ces structures formelles ont été produites à partir de la page Map Features d'OSM.

D'une part, l'ontologie OSMOnto (Codescu et al. 2011) a été manuellement dérivée du site de documentation (en anglais) des tags Map Features dont les tags considérés sont ceux les plus

[44] La logique formelle de Horn, en calcul propositionnel, permet d'exprimer des règles, faits et requêtes. La procédure d'inférence calcule les réponses à une requête à partir de ces règles et faits. Toute clause de Horn est exprimée de la forme : $r_1 \wedge r_2 \wedge \ldots \wedge r_n \rightarrow h$

fréquemment utilisés par les contributeurs selon l'outil Taginfo. Un des objectifs d'OSMonto est d'organiser les tags OSM dans une structure taxonomique afin de faciliter le traitement par une machine du modèle de tags OSM. Dans OSM, un tag est encodé comme `clé=valeur`$_1$, `clé=valeur`$_2$, ..., `clé=valeur`$_N$ (voir la section I.3.2). Dans OSMonto, cette `clé` du tag représente une classe `k_clé` et puis chacune de ses `valeurs` (`v_valeur`$_1$, `v_valeur`$_2$, ..., `v_valeur`$_N$) lui est associée comme une superclasse (via une relation `is-a`). Plus précisément, cette ontologie possède une structure taxonomique à deux niveaux avec une racine correspondant à la classe `Thing`. Le niveau #2 correspond à des classes associées aux clés des tags OSM, c'est-à-dire, les classes de type `k_clé`. Une classe niveau #2 corresponde à un thème géographique (ex : utilisation du sol, commerces, etc.). Le niveau #1 correspond à des classes associés aux valeurs des tags OSM, c'est-à-dire, les classes de type `v_valeur`$_i$. Une classe niveau #1 décrit précisément la nature d'une entité géographique dans OSM et correspond à un thème géographique (ex : la classe `boulangerie` du thème « commerce »). OSMonto contient une totalité de 471 classes, 29 types de relation (*Object Properties*), 27 types de propriétés (*Data Properties*).

D'autre part, le graphe OSM Semantic Network (Ballatore et al. 2012) a été dérivé automatiquement à partir d'un robot d'indexation (*Web Crawler*). Il explore automatiquement les pages Web décrivant les tags OSM[45] et leurs liens hypertextes dans le site Map Features d'OSM. La Figure 28 montre la description en RDF de la ressource d'OSM Semantic Network correspondante au tag `building = hall`. Sachant que les nouveaux tags sont régulièrement documentés sur le site de Map Features, l'avantage d'un robot d'indexation est de permettre la récupération automatique des nouvelles versions de la documentation des tags sous une forme d'un graphe. Ainsi, il est plus facile pour un contributeur, d'explorer les tags OSM et de trouver des incohérences dans le modèle. De plus, l'avantage de ce graphe est qu'il capture bien plus des relations décrites entre les tags sur le site de Map Features. En revanche, l'ontologie OSMonto contient principalement des liens *is-a* et est limitée à deux niveaux.

Baglatzi et al. (2012) proposent d'aider les contributeurs à aligner le système de tags OSM avec des ontologies de référence existantes, plus précisément l'ontologie de haut niveau DOLCE (Gangemi et al. 2002) [46], d'une façon ludique. Grâce à l'alignement des tags avec une ressource formelle, il est possible de formaliser la représentation des tags en explicitant par exemple, des relations `is-a`. Ainsi, un raisonneur peut détecter automatiquement certaines incohérences. Par exemple, certaines valeurs du tag `amenity` correspondent à des descriptions d'un PAI et d'autres valeurs de ce tag à des activités que les gens peuvent réaliser dans le PAI. L'alignement avec DOLCE permet d'expliciter ces distinctions entre la description et l'activité pour un PAI.

[45] Il existe une page web sur le site OSM pour décrire exhaustivement chaque tag, voir par exemple la page décrivant le tag `library` : http://wiki.openstreetmap.org/wiki/Tag:amenity%3Dlibrary

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

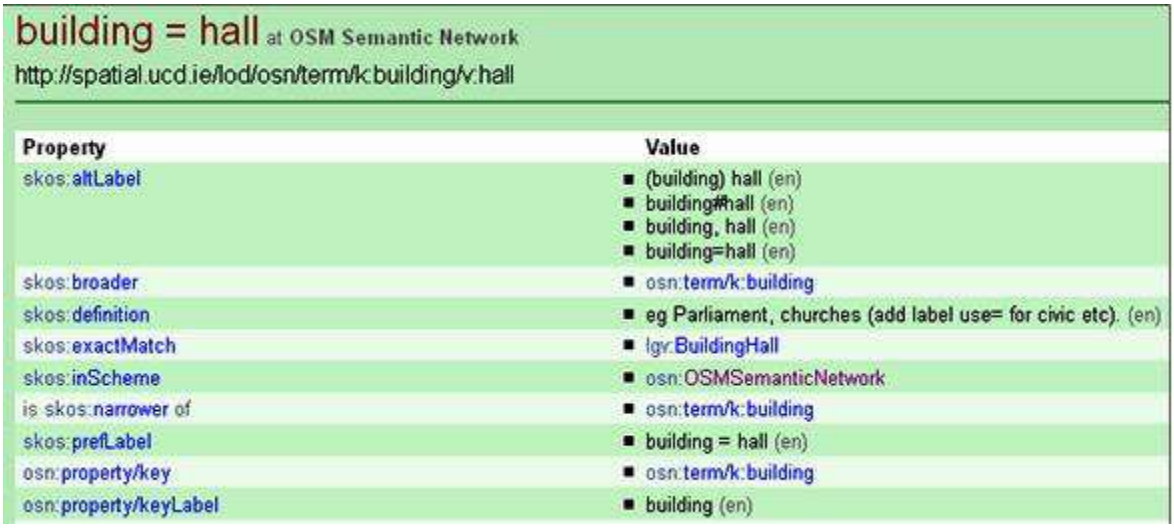



**Figure 28 : description en RDF de la ressource de correspondante au tag building = hall**

## 2.2 Les méthodes d'évaluation de la qualité des données

Étant donnée la nature des contenus dits VGI, les méthodes d'évaluation de la qualité de ces données varient de celles des producteurs nationaux (Goodchild 2009). Des recherches effectuent des comparaisons entre des données OSM et des données de référence issues d'un producteur traditionnel (public ou privé) de données géographiques. D'autres travaux sont à la recherche des nouvelles méthodes afin de qualifier les contributions en prenant en compte le profil du contributeur et l'activité d'édition encapsulée dans un historique des données. Plus récemment, certains travaux s'intéressent à la visualisation de la qualité du VGI.

### 2.2.1 Comparaison suivant la norme ISO 19115 avec des données de référence

Des études d'évaluation de la qualité de données OSM ont été effectués sur certaines villes européennes en comparant ces données avec des données de référence issues d'un producteur traditionnel de données géographiques (public ou privé). Ces études se basent sur les métriques classiques de qualité dictées par la norme ISO 19115 (ISO 2003) : précision géométrique, précision attributaire et sémantique, exhaustivité, cohérence logique, actualité et origine. Haklay (2010) mesure la précision géométrique et l'exhaustivité du réseau routier OSM datant de 2008 sur la ville de Londres et d'autres villes anglaises en utilisant la méthode d'appariement de Goodchild et Hunter (1997) basée sur des *buffers*. Cette méthode permet de déterminer le pourcentage de géométries linéaires d'un objet appartenant au jeu de données OSM et se trouvant dans l'aire du *buffer* des géométries linéaires de l'objet dans le jeu de données Ordnance Survey. Zielstra et Zipf (2010) utilisent la même méthode pour évaluer la qualité des données OSM correspondantes à des villes importantes et de taille moyenne en Allemagne datant de 2009 avec des données TeleAtlas. Girres et Touya (2010) effectuent une évaluation de qualité de données OSM datant de 2009 en France avec des données de l'IGN. A la différence des études précédentes, ils considèrent l'ensemble entier de métriques proposées par la norme ISO 19115 et utilisent d'autres méthodes d'appariement basées sur la distance de Hausdorff pour comparer des géométries linéaires et la distance de Vauglin (1997) pour comparer des géométries surfaciques. Ces études montrent globalement la faible exhaustivité de ces données dans les zones rurales par rapport à des zones urbaines. Girres et Touya (2010) signalent des incohérences dans les données à cause de l'absence de spécifications

précises (non ambigües), voire formelles. En effet, certains types d'objets, comme les routes, ont été mieux classifiés que d'autres par les contributeurs grâce à la présence d'une classification plus claire et exhaustive disponible dans le site de Map Features pour ces types d'objets. Ces auteurs identifient également des incohérences concernant l'absence du partage de la géométrie. La Figure 29 montre des incohérences entre des limites administratives (en vert), et entre des limites administratives et des lignes côtières (Girres et Touya 2010).

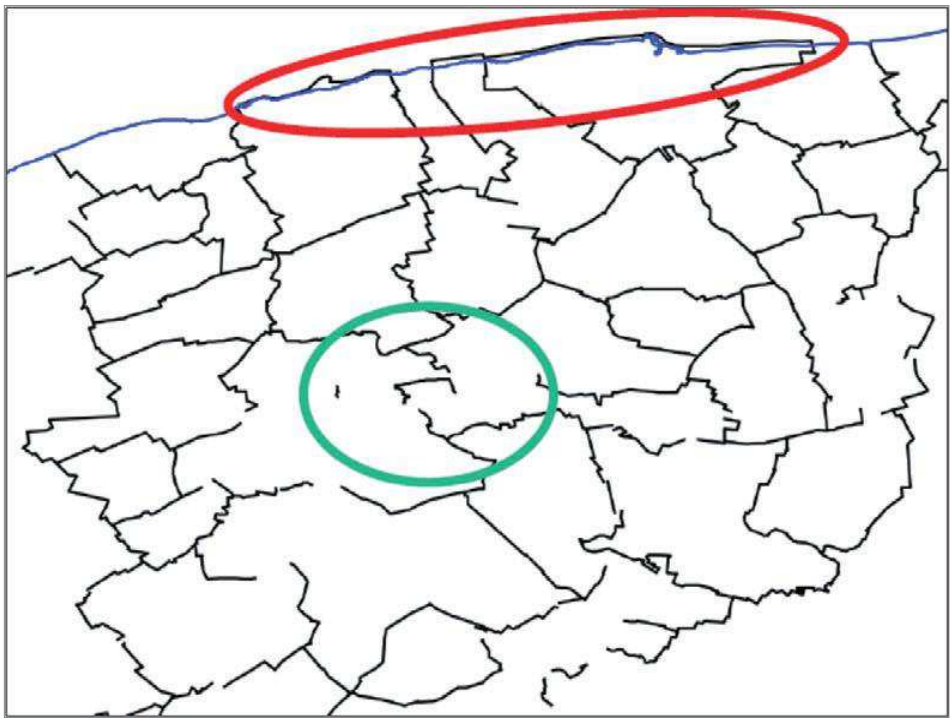


**Figure 29 : exemple de (Girres et Touya 2010) montrant des incohérences pour le département de la Seine-Maritime entre des limites administratives (en vert) et l'absence du partage des géométries entre des lignes côtières et des limites administratives (en rouge)**

De plus, ces travaux étudient la relation entre le nombre de contributeurs et la qualité des données OSM, inspirés par la Loi de Linus suivante : « avec suffisamment d'yeux, les bugs sont minimisés »[47]. Pour la précision géométrique, l'écart entre les données OSM et les données de référence est moins important dans les zones possédant plus de 15 contributeurs et moins de 5 contributeurs Haklay et Basiouka (2010). Pour l'exhaustivité, cette relation augmente progressivement mais pas d'une façon linéaire (Girres et Touya 2010).

### 2.2.2 Qualification des contributions et des contributeurs

Les méthodes classiques d'évaluation de la qualité de données géographiques ne prennent pas en compte la dimension collaborative du VGI fortement concernée par l'hétérogénéité des contributeurs et la considérable activité collaborative d'édition des données. Pour cette raison, les recherches actuelles s'intéressent à concevoir de nouvelles méthodes pour qualifier les contributions en prenant en compte les contributeurs et leurs activités d'édition des données (par l'historique de données).

[47] La loi de Linus décrit la philosophie de développement de logiciel libre (Raymond 1999).

Coleman et al. (2009) et Budhathoki (2010) s'intéressent à l'étude des motivations des contributeurs afin de qualifier implicitement leurs contributions. En effet, la motivation des contributeurs est fortement reliée à la crédibilité du VGI (Flanagin et Metzger 2008). Par exemple, la motivation d'un contributeur peut être de vouloir soutenir une communauté en ligne. En analysant les résultats de recherche concernant les motivations des contributeurs de Wikipédia et des logiciels open source, Coleman et al. (2009) esquisse une liste de critères qui peuvent motiver un contributeur du VGI : altruisme, intérêt personnel ou professionnel, stimulation intellectuelle, amélioration et protection du projet communautaire, gratification sociale, amélioration de sa réputation entre les pairs de la communauté, expression créative, fierté de sa localité (ex : voir son village bien cartographié sur une carte), et vandalisme (ex : effacer un grand lot de données ou inventer des attributs). Afin d'identifier empiriquement les motivations des contributeurs OSM, Budhathoki (2010) analyse environ 3000 conversations entre contributeurs sur les listes de diffusions, de même que des contributions de 34.000 contributeurs (entre 2004 et 2009) et propose un sondage à des contributeurs OSM. L'auteur trouve que les contributeurs sont globalement concernés par les zones vides (sans aucune contribution), par des zones contenant des erreurs, et par des zones qui leurs sont familières. Les participants au sondage signalent également que l'altruisme n'est pas une motivation. Néanmoins d'après cette étude, la relation exacte entre les motivations du contributeur et la qualité du contenu n'est pas encore claire.

Bishr et Kuhn (2007) et Bishr et Janowicz (2010) conçoivent un modèle spatio-temporel fondé sur des critères de confiance et de réputation afin de qualifier les contributions de VGI. Les auteurs s'inspirent des travaux sur les réseaux sociaux. Ils expliquent que « si un contributeur A peut signaler un autre contributeur B comme quelqu'un ayant une bonne réputation, alors il est possible de faire confiance à la contribution de B ». C'est-à-dire qu'un contributeur de confiance fournit en général des informations plus pertinentes qu'un contributeur de moindre confiance. Leur modèle considère qu'une bonne contribution est celle qui a été reportée plusieurs fois par des contributeurs différents. Par exemple, considérons deux contributeurs A et B qui fournissent séparément deux observations correspondantes à une même réalité : un nouveau bâtiment a été construit dans tel endroit. Le système calcule la proximité entre les objets saisies par A et B, et détermine que les deux nouveaux bâtiments sont probablement un même et unique bâtiment. Ensuite, le système informe A et B de cette conclusion. Si les contributeurs sont d'accord, les contributions sont fusionnées et la contribution résultante est signalée comme la même observation reportée par deux contributeurs différents. Les autres contributeurs peuvent donc faire confiance à cette observation. Leur modèle envisage également le cas où deux observations sur le même phénomène réalisées par deux individus sont contradictoires : le système doit alors vérifier les intervalles de temps des contributions. Si l'intervalle de temps est large, alors il est probable que le changement se soit passé dans la réalité. Sinon, il n'est pas probable qu'un changement si soudain se soit passé. Dans ce cas, ces deux observations peuvent donc être considérées comme des « mauvaises » contributions.

D'autres travaux s'intéressent à l'historique des données OSM pour l'analyse de l'activité d'édition afin de qualifier les contributeurs et leurs contributions. Keßler et al. (2011) s'intéresse particulièrement à déterminer automatiquement la réputation d'un contributeur, et ainsi la confiance en sa contribution, par des règles en utilisant cet historique. Les auteurs proposent un formalisme de logique de Horn pour expliciter les éditions sur les données OSM (ex : `removesTag`, `addsTag`, `changesValueofKey` et `changesGeometry`). Ensuite, il est possible de définir des règles servant à détecter, dans l'historique des données, des comportements d'édition de type `correction` (ex : un contributeur corrige une donnée), `confirmation` (ex : un contributeur confirme les informations existantes sur une donnée), et

`annulation` (ex : un contributeur revient un changement en arrière). L'Équation 2 montre un exemple d'une règle pour diminuer la réputation d'un contributeur A si sa contribution a été tout de suite corrigée par un contributeur B (c'est-à-dire, une `correction`). En conséquence, la réputation de A sera diminuée.

$$FeatureState(?f1) \wedge FeatureState(?f2) \wedge precededBy(?f2,?f1) \wedge Edit(?e1) \wedge Edit(?e2) \wedge createdBy(?f1,?e1) \wedge createdBy(?f2,?e2) \wedge changesValueOfKey(e2,?t) \rightarrow Correction(?e2)$$

**Équation 2 : expression en logique de Horn décrivant une règle de correction quand un contributeur corrige tout de suite la contribution d'un autre (Keßler et al. 2011).**

### 2.2.3 La visualisation de la qualité

Peu de recherches en VGI s'intéressent à l'usage de solutions visuelles afin de permettre aux utilisateurs d'observer les forces et les faiblesses des données. Dans OSM, Roick et al. (2011) proposent OSMatrix[48], une application Web permettant la visualisation de plusieurs indicateurs concernant le comportement des contributeurs, l'actualité, l'exhaustivité relative des données, et l'agrégation de plusieurs types d'objets (ex : les zones commerciales, résidentielles, etc.). OSMatrix permet de connaître par exemple, le nombre d'objets modifiés par contributeur, le nombre de versions, les dates de dernières modifications, le nombre d'objets et de leurs tags correspondants. La Figure 30 montre le centre de la ville de Paris et la visualisation du nombre maximum d'attributs associés aux objets dans la cellule (`id = 628147`) où le numéro maximum d'attributs est 15.

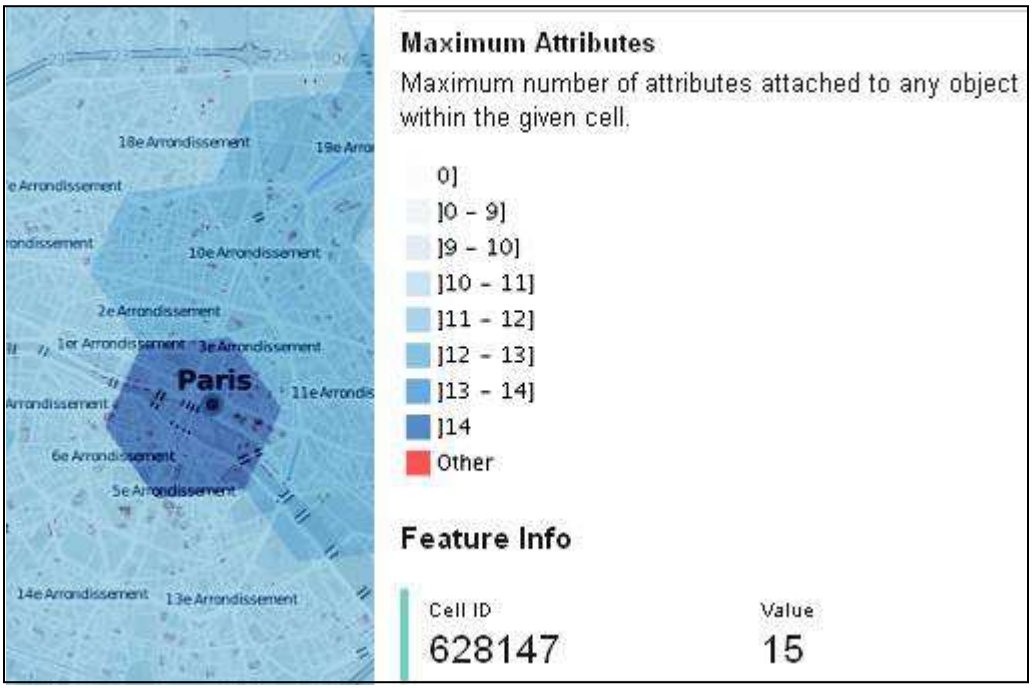



**Figure 30 : visualisation sur l'application Web OSMatrix (Roick et al. 2011) du nombre maximum d'attributs associé aux objets dans plusieurs dalles de Paris**

[48] http://koenigstuhl.geog.uni-heidelberg.de/osmatrix/

# 3 La gestion de la cohérence de données géographiques

Dans cette section, nous décrivons quelques recherches en sciences de l'information géographique visant à améliorer la cohérence des données géographiques pour la gestion de leur qualité. Étant donnée la grande quantité de travaux existants, nous ne pouvons pas être exhaustifs et nous nous concentrons sur certains éléments qui semblent pertinents pour gérer la cohérence d'un contenu géographique dans un contexte d'édition collaborative.

## 3.1 Les contraintes d'intégrité pour des données géographiques : le rôle des relations spatiales

Cette section décrit des travaux sur la définition de contraintes d'intégrité pour des données géographiques pour la gestion automatique de la cohérence des données géographiques. Ces contraintes d'intégrités sont fréquemment définies sous forme de relations spatiales qui doivent être préservées dans le contenu (Egenhofer et Mark 1995; Borges et al. 2002; Duchêne 2004; Bejaoui et al. 2010).

### 3.1.1 Les relations spatiales

Les relations spatiales sont primordiales pour décrire la configuration de l'espace géographique (Bruns et Egenhofer 1996). Ces relations peuvent être dérivées des géométries d'objets par des opérations d'analyse spatiale (appelées des relations spatiales implicites). Dans un autre cas, ces relations peuvent être définies explicitement (appelées des relations spatiales explicites) dans le schéma de données de l'application qui est en charge de garder les liens explicites (Hadzilacos et Tryfona 1992; Borges et al. 2002). Il existe trois types de relations. Les relations topologiques sont inhérentes au concept de la connectivité et sont invariantes aux transformations topologiques comme la rotation, la translation et le changement d'échelle (ex : `contient`, `à l'intérieur` et `croise`). Les relations d'orientation supposent l'existence d'un axe de référence et sont variantes aux transformations topologiques (ex : `à droite de`, `à gauche de`, `derrière`, `en face`, `au nord`, `au sud`). Les relations métriques expriment le concept de distance et varient en fonction du changement d'échelle mais non à la rotation et translation (ex : `près de`, `loin de`). Les domaines de valeurs d'un type de relations sont restreints à certains types de géométries (points, lignes, et polygones) des classes impliquées.

### 3.1.2 Les contraintes d'intégrité

L'utilisation de contraintes d'intégrité de la cohérence des données pendant, soit la mise à jour de la base de données, soit l'acquisition des données, ou soit la généralisation cartographique (Laurini et Thompson 1992; Cockcroft 1997; Servigne et al. 2000; Duchêne 2004). Ces contraintes d'intégrité sont fréquemment définies sous la forme de relations spatiales qui doivent à être préservées dans le contenu.

Les classifications des contraintes d'intégrité existantes dans la littérature partent de la classification proposée par Elmasri et Navathe (1994) dans une approche des bases de données relationnelles. En partant de cette classification et en considérant les particularités des données géographiques, Cockcroft (1997) propose une classification des contraintes d'intégrité

en trois groupes. Les contraintes topologiques portent sur la topologie des objets, par exemple « les composantes d'une partition sont toutes disjointes ». Les contraintes dites « sémantiques » portent sur la nature même des objets dans le monde réel, par exemple « un bâtiment ne chevauche pas le tronçon de route ». Les contraintes dites « utilisateur » sont définies selon les besoins spécifiques d'un utilisateur, par exemple « une station d'essence pour des raisons de sécurité devrait être à 200 mètres minimum d'une école ». Plus tard, Mäs et Reinhardt (2009) adaptent la classification précédente en proposant des contraintes d'intégrité basées sur la typologie de relations spatiales proposée par Mark et Frank (1989) et Egenhofer et Franzosa (1991). Ainsi, Mäs et Reinhardt (2009) proposent les types de contraintes d'intégrité suivants : les contraintes topologiques (ex : « les lacs ne croisent pas les courbes de niveau »), d'orientation (ex : « l'arrière-cour doit être derrière la maison »), métriques (ex : « une station d'essence doit être à une distance minimum de 200 mètres d'une école »). De plus, les auteurs proposent des contraintes complexes combinant des relations précédemment mentionnées et des relations d'agrégation, par exemple « le nombre d'habitants d'un pays est la somme du nombre d'habitants de ses régions administratives ».

Des formalismes ont été proposés afin de définir des contraintes d'intégrité sur des types de relations spatiales, ainsi que des méthodologies pour évaluer automatiquement ces contraintes sur les données. Leur objectif est d'assurer la cohérence logique des données pendant leur acquisition. En effet, la validation automatique de règles aide à minimiser le temps d'acquisition des données et à s'assurer de leur qualité (Mäs et al. 2005). Ces travaux seront décrits ci-après.

Servigne et al. (2000) proposent une méthodologie pour évaluer la cohérence d'une base de données géographique existante. Globalement, leur objectif est de fournir des procédures de validation pendant l'acquisition des données afin de préserver la cohérence de la base de données. Pour définir les contraintes d'intégrité, l'utilisateur définit des contraintes via une interface graphique en utilisant des types de relations topologiques prédéfinis. Ces contraintes concernent le respect de la relation topologique entre deux objets en prenant en compte leur nature dans le monde réel. Par exemple, un bâtiment est rarement à l'intérieur d'un autre bâtiment, mais un bâtiment peut être à l'intérieur d'une parcelle. Dans ces cas, le type de relation `est à l'intérieur`, est défini entre deux types de géométries polygonales. Les auteurs utilisent le modèle de neuf intersections (9IM) (Egenhofer et Herring 1991) permettant de formaliser et d'identifier les relations topologiques entre deux objets avec des géométries simples (point, ligne, surfacique). Chaque contrainte est représentée comme `C(EntityClass1, Relation, EntityClass2, specification)`, par exemple la « contrainte le route ne doit pas être à l'intérieur du bâtiment » est instanciée ainsi `C1(Road, Inside, Building, Forbidden)` par l'utilisateur via l'interface graphique. D'autres exemples sont `C2(Road, Cross, Building, Forbidden)` et `C3(Sluice, Joint, Waterpipe, Exactly 2 times)`.

Ensuite, en ce qui concerne l'évaluation de ces contraintes, chacune est traduite comme une conjonction des relations spécifiées dans la contrainte. La relation entre deux objets est calculée en utilisant la matrice 9IM afin de trouver les intersections existantes entre les frontières, les intérieurs et les extérieurs des deux objets. Dans le cas d'erreurs potentielles, le système peut calculer plusieurs scénarios de correction et les suggérer à l'utilisateur via l'interface graphique. Par exemple, la manière de corriger la violation d'une contrainte concernant une relation topologique est de changer la relation entre les deux objets en appliquant les transformations suivantes : `déplacer`, `changer la forme`, `effacer`, ou `découper les objets`. La transformation `déplacer` consiste de déplacer l'objet vers une

direction arbitraire. Le calcul de cette transformation sur plusieurs directions (nord, sud, est, ouest) permet d'avoir plusieurs scénarios possibles. Ces corrections sont proposées à l'utilisateur. L'utilisateur peut décider qu'il n'y a pas d'incohérence mais plutôt une exception, il peut accepter une correction ou peut effectuer lui-même une correction. Le principal avantage de calculer et de proposer plusieurs corrections est de faciliter et accélérer le travail de l'utilisateur, mais aussi de contrôler le processus de correction afin de ne pas créer de nouvelles erreurs. Néanmoins, la façon dont de telles corrections sont calculées n'est pas spécifiée par les auteurs.

Mäs et al. (2005) proposent la définition formelle des contraintes d'intégrité dans le langage de définition de règles du Web sémantique (SWRL) qui est un candidat au standard W3C depuis 2004. SWRL est l'union du langage de formalisation des ontologies *Web Ontology Language* (OWL) et le langage *Rule Markup Language* (RML) pour la définition de règles de Horn (`prémisse → conséquence`). Les auteurs adoptent la proposition de Frank (2001) de formaliser ces contraintes en étendant une ontologie. Une contrainte est définie comme une règle en logique de Horn, c'est-à-dire, si la prémisse est vraie alors la conséquence doit aussi être vraie. La prémisse et la conséquence sont des conjonctions d'axiomes faisant référence à des concepts de l'ontologie (Horrocks et al. 2004). Une contrainte possède aussi un identifiant unique, un dégrée d'importance, une description de la contrainte et une instruction de correction en langage naturel. Le degré d'importance permet de classer les traitements dans un cas de violation de plusieurs contraintes. La valeur `strict` indique que les données doivent être changées en suivant le conseil proposé. Les valeurs `éviter la violation` et `avertissement : l'intervention de l'utilisateur est nécessaire` délèguent à l'utilisateur la décision sur la correction à appliquer. La troisième valeur force l'intervention de l'utilisateur. En revanche, son intervention est optionnelle pour la deuxième valeur. Par exemple, l'Équation 3 montre la contrainte topologique « les routes ne doivent pas croiser un fossé » en SWRL. Il faut clarifier que les auteurs ne précisent pas la méthodologie pour évaluer ces contraintes sur les données.

```
<rule:imp>
    <swrlagis:constrainteID> 1 </swrlagis:severity>
    <swrlagis:gravite> strict </swrlagis:gravite>
    <swrlagis:correction> Couper la route en deux tronçons
    </swrlagis:correction>
    <swrlagis:commentaire> les routes ne doivent pas s'intersecter
    un fossé </swrlagis:commentaire>
     <ruleml:_body>
          <swrlx:classAtom>
              <owlx:Class owlx:name="Way"/>
              <ruleml: var>way</ruleml:var>
         </swrlx:classAtom>
         <swrlx:classAtom>
              <owlx:Class owlx:name="Ditch"/>
              <ruleml: var> ditch </ruleml:var>
         </swrlx:classAtom>
     </ruleml:_body>
     <ruleml:_head>
         <swrlx:individualPropertyAtom swrlx:property="intersect">
              <ruleml:var>way</ruleml:var>
              <ruleml:var>ditch</ruleml:var>
```

```
        </swrlx:individualPropertyAtom>
    </ruleml:_head>
</rule:imp>
```

**Équation 3 : une contrainte topologique « les routes ne doivent pas croiser un fossé » s'exprime en SWRL (simplifié) (Mäs et al. 2005)**

Pinet et al. (2009) utilisent le langage formel déclaratif OCL (*Object Constraint Language*) et proposent l'extension $OCL_{9IM.}$ Cette extension permet de modéliser des contraintes portant sur des relations topologiques entre des objets surfaciques simples et composites. Cette extension est basée sur le modèle de 9IM (Egenhofer et Herring 1991) permettant de formaliser et identifier les relations topologiques entre deux objets avec des géométries simples (ponctuelles, linéaires, surfaciques). Une contrainte d'intégrité indiquant soit `I` un îlot et `C` un centre-ville, si les contraintes `I` et `C` sont associées par `centre_ville_contenant_les_îlots` alors pour chaque bâtiment `b` de `I`, il doit exister une partie `c` du centre-ville `C` tel que `b` est dans `c`. L'Équation 4 exprime la contrainte en $OCL_{9IM}$.

```
context Ilot inv:
   self.geo -> forAll( b |
     self.centre_ville_contenant_les_ilots.geo
        -> exists ( c | (b). inside (c)))
```

**Équation 4 : contrainte en $OCL_{9IM}$ pour vérifier qu'il n'existe pas de cas où une parcelle d'épandage et sa commune principale sont disjointes (Pinet et al. 2009)**

Les auteurs proposent aussi une deuxième extension appelée $OCL_{ADV}$ qui étend $OCL_{9IM}$ et qui possède la même expressivité. Cette extension facilite l'écriture de contraintes car il est possible d'associer à une relation, un adverbe (Claramunt 2000) : `occasionnellement`, `jamais`, `entièrement`. L'Équation 5 retranscrit en $OCL_{ADV}$ la contrainte précédemment indiquée dans l'Équation 4.

```
context Ilot inv:
   (self.geo) -> inside
   ("mostlyRev", self.centre_ville_contenant-les_ilots.geo)
```

**Équation 5 : contrainte en $OCL_{ADV}$ pour vérifier que des bâtiments sont bien à l'intérieur de leur ilot correspondant (Pinet et al. 2009)**

Pour l'évaluation des contraintes d'intégrité sur des données, les auteurs proposent un générateur de code OCL2SQL qui traduit automatiquement la contrainte en SQL sous la forme d'une procédure de type déclencheur (*trigger*) dans le système de gestion de base de données Oracle. La traduction est effectuée à partir d'un fichier XML décrivant le diagramme de classe de l'application, d'un fichier de métadonnées relatives aux attributs géographiques et des contraintes spatiales en $OCL_{ADV}$ ou $OCL_{9IM}$. Ensuite, ces déclencheurs sont exécutés au moment de la mise à jour de la base de données. Par exemple, l'Équation 6 montre la contrainte en $OCL_{9IM}$ exprimant : « la configuration topologique disjoint n'est pas tolérée entre une parcelle d'épandage et sa commune principale, dans le contexte de la gestion de propositions d'épandage agricole de matière organique en France » (Pinet et al. 2009). Ensuite, l'Équation 7 montre la traduction de cette contrainte en SQL.

```
context Parcelle inv:
   (not) ( (self.geo). disjoint (self.commune_ratachée.geo) )
```

**Équation 6 : contrainte en $OCL_{9IM}$ pour vérifier que la disjonction entre une parcelle d'épandage et sa commune principale n'est pas tolérée (Pinet et al. 2009)**

```
SELECT * PARCELLE SELF WHERE NOT
  (NOT (
    MDSYS.SDO_RELATE (
      SELECT GEO FROM COMMUNE WHERE ID_COMMUNE IN
        (SELECT COMMUNE_RATACHEE_ID_COMMUNE FROM PARCELLE
         WHERE ID_PARCELLE = PARCELLE.ID_PARCELLE))
         , SELF.GEO
         , 'mask = DISJOINT   querytype=WINDOW') = 'TRUE')));
```

**Équation 7 : traduction en SQL de la contrainte OCL9IM montrée en Équation 6 (Pinet et al. 2009)**

Par la suite, cette commande sera exécutée pendant la mise à jour de la base de données afin de valider leur cohérence par rapport à ces règles.

Werder (2009) propose également une extension géométrique d'OCL baptisé GeOCL. Ils étendent la grammaire d'OCL pour ajouter le type de donnée Géométrie. L'auteur fournit également des opérations nécessaires afin de manipuler des instances de Géométrie (ex : calcule de la superficie, intersection) durant l'évaluation des contraintes sur des géométries. Par exemple, il est possible de définir et évaluer la contrainte « les murs antibruit son disjoints des géométries des routes » exprimée en GeOCL dans l'Équation 8. Cette contrainte est évaluée en appliquant l'opération `intersection` sur les géométries des objets représentant les murs antibruit et les routes.

```
context NoiseAbatementWall
inv: Streets.allInstances() ->
      forAll (s:Street|s.geometry -> disjoint(self.geometry))
```

**Équation 8 : contrainte en GeOCL exprimant la contrainte « les murs anti-bruits son disjoints des géométries des routes « (Werder 2009)**

## 3.2 L'appariement et la transformation des données géographiques

Cette section présente une vue générale des travaux sur l'appariement (*data matching*) et la transformation (*feature transformation*) de plusieurs jeux de données géographiques (vecteurs ou images) venus de sources hétérogènes. Ces processus sont groupés dans la littérature sous le terme *conflation* (Saalfeld 1993; Chen et Knoblock 2008), aussi appelée par d'autres auteurs l'intégration de données géographiques (Devogele et al. 1998; Sheeren et al. 2004). L'objectif de la *conflation* des données vecteurs est de produire à partir de plusieurs jeux de données, un nouveau jeu de données avec une "meilleure" précision spatiale et attributaire, en minimisant la redondance et en corrigeant les incohérences des données (Longley et al. 2005). Quand la *conflation* est faite à partir de données sur une même zone, on parle de *conflation* verticale, et quand elle est faite sur des zones limitrophes, on parle de *conflation* horizontale. La *conflation*

peut être utilisée pour l'intégration de données et la gestion de la qualité pendant leur intégration (Bel Hadj Ali 2001; Stankutė et Asche 2011), de même que pour la mise à jour (Badard 2000), et l'enrichissement des données (Walter et Fritsch 1999). La Figure 31 présente la classification des processus de *conflation* trouvée dans la littérature.

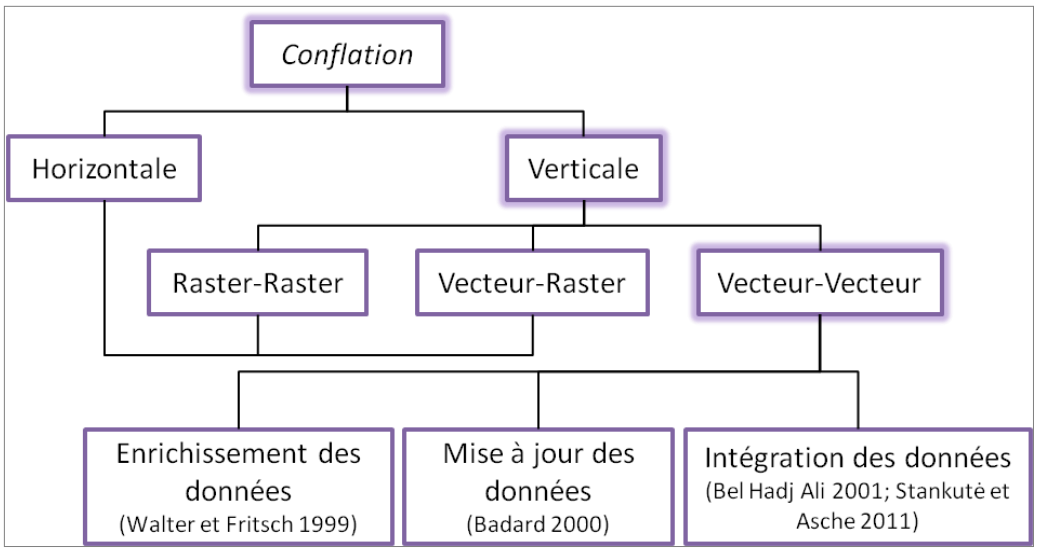



**Figure 31 : la classification (simplifiée) des processus de conflation trouvée dans la littérature (Yuan et Tao 1999)**

La *conflation* se décompose en plusieurs processus, comme illustrés en Figure 32. La phase de prétraitement des données s'assure que les deux jeux de données ont le même format, échelle, projection cartographique, et système de référence. Nous décrivons ensuite les phases d'appariement et de transformation.

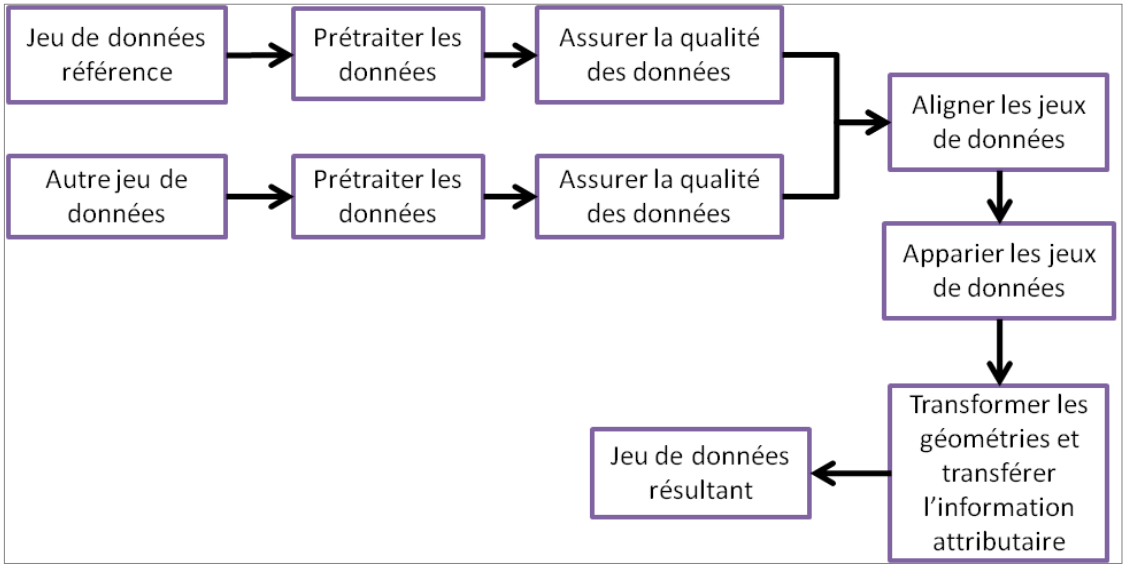



**Figure 32 : la décomposition des processus de conflation (Yuan et Tao 1999)**

### 3.2.1 L'appariement

La phase d'appariement vise à déterminer les éléments homologues de deux jeux de données en s'appuyant sur des filtres sur les géométries (ex : la distance la plus courte), les relations topologiques (ex : la similarité vis-à-vis les relations topologiques), et la « sémantique » des objets afin de choisir les meilleures correspondances (Yuan et Tao 1999; Casado 2006; Ruiz et al. 2011). Dans plusieurs cas, il est nécessaire d'avoir une phase d'évaluation et de validation des correspondances trouvées.

Les approches qui prennent en compte les géométries utilisent des distances absolues comme Hausdorff (Deng et al. 2005), et Fréchet (Devogele 2002) pour mesurer la similarité des objets linéaires. Bel Hadj Ali (2001) s'intéresse à la similarité entre des objets surfaciques complexes (ex : polygones à trous et agrégats de polygones). Il propose une méthode d'appariement qui prend en compte les contours et les intérieurs des entités. Plus précisément, la méthode utilise la distance de Hausdorff entre les contours des polygones (Abbas 1994), la distance entre fonctions angulaires pour évaluer la différence de forme entre deux polygones (Arkin et al. 1991), et la distance surfacique entre les aires des polygones (Vauglin 1997). Dans sa thèse, Bel Hadj Ali (2001) (p. 86 et 92) rappelle les définitions de ces mesures de la manière suivante :

- *la distance de Hausdorff entre deux contours C1 et C2 est le maximum de deux quantités, la première est le maximum des plus courtes distances (généralement, la distance euclidienne) des points du contour C1 à l'ensemble des points du contour C2, et la seconde est le maximum des plus courtes distances euclidiennes de l'ensemble des points du contour C2 à l'ensemble des points du contour C1,*
- *la distance entre fonctions angulaires utilisée pour comparer les formes des polygones en se basant sur leurs fonctions angulaires. Une fonction angulaire décrit l'entité surfacique à travers les angles formés par les segments qui composent son contour, et*
- *la distance surfacique entre deux polygones A et B, est le rapport de l'aire de la différence symétrique de A et B, et l'aire de l'union de A et B.*

L'appariement dit « sémantique » utilise les ontologies pour évaluer la similarité entre les objets en considérant les relations sémantiques existantes (Duckham et al. 2006; Ressler et al. 2009). Abadie (2012) propose un modèle fondé sur des ontologies OWL2 pour la formalisation des connaissances issues des spécifications de bases de données géographiques pour guider leur appariement. Cet auteur affirme qu'il est nécessaire d'inclure l'information sur la façon dont les géométries ont été acquises afin de guider l'appariement des données.

Les approches d'appariement multicritères considèrent les informations sur les géométries, la topologie, et l'information attributaire pour choisir les meilleurs candidats (Cobb et al. 1998; Mustière et Devogele 2008; Adams et al. 2010; Li et Goodchild 2011). Plus particulièrement, Olteanu (2008) représente explicitement des connaissances et leurs imperfections pour définir les critères d'appariement. Cet auteur utilise la théorie des fonctions de croyance de Dempster-Shafer (Dempster 1967; Shafer 1976) afin de modéliser les possibles imperfections existantes dans les données : l'imprécision, l'incertitude et incomplétude. En effet, certaines données géographiques sont imprécises ou peuvent avoir des erreurs introduites par l'opérateur ou par le logiciel de saisie. Dans sa thèse, Olteanu (2008) (p. 79) rappelle ces trois définitions, composant la taxonomie d'imperfections, de la manière suivante :

- *Imprécision : concerne la difficulté d'exprimer clairement et précisément un état de la réalité par une proposition (par exemple « dans la salle il y a environ une centaine de*

*personnes », ou « le poids de la table est d'environ 25 kg », ou « Jean est grand ». Modéliser l'imprécision consiste à formaliser les termes de « environ », « centaine » ou « grand »,*

- *Incertitude : concerne un doute sur la validité d'une connaissance. Elle est due à la fiabilité de l'observateur peu sûr de lui ou prudent qui ne peut pas déterminer la valeur de vérité de la connaissance. Ex : « Je crois que dans la salle il y a 100 personnes »,*
- *Incomplétude : il s'agit d'une absence de connaissance ou d'une connaissance partielle. Elle est due à une incomplétude dans les données, à l'absence d'une connaissance explicite ou à l'existence d'une connaissance générale. Par exemple, pour une instance d'une base de données la valeur de l'attribut Nom n'est pas remplie.*

La méthode d'appariement d'Olteanu (2008) prend en compte trois types de connaissances :

- celles issues des données géographiques qui permettent de calculer les mesures de distance entre les objets géographiques des deux jeux de données,
- celles issues des spécifications des bases de données géographiques qui sont utilisées pour définir les seuils,
- et enfin les connaissances issues des experts qui nous permettent de définir les masses de croyance.

Plus précisément, cet auteur propose une méthodologie composée de cinq étapes principales : la sélection des candidats, l'initialisation des masses de croyance, la fusion des critères d'appariement pour chaque candidat, la fusion des candidats à l'appariement et la décision. Les fonctions de croyance par candidat du jeu de données d'évaluation, sont calculées selon deux critères : géométrique, par la distance Euclidienne entre les coordonnées géographiques des entités, et toponymique, par la distance entre les deux chaînes de caractères. Les critères sont fusionnés grâce à l'opérateur de (Dempster 1967). Le processus d'appariement proposé par Olteanu (2008) peut être appliqué d'une part aux trois types de représentation (points, lignes ou surfaces) et aux différents thèmes (routier, hydrographique, bâtiments, occupation du sol, etc.) et d'autre part aux jeux de données ayant le même niveau de détail ou des niveaux de détail différents.

### 3.2.2 La transformation

La phase de transformation des objets effectue globalement des opérations géométriques et des transferts d'attributs (Yuan et Tao 1999; Wiemann et Bernard 2010; Ruiz et al. 2011) entre une entité d'un jeu de données et son entité homologue de l'autre jeu de données. Par exemple, si deux objets représentant la même entité du monde réel ont des valeurs différentes sur leurs attributs du même nom, une règle admise est de prendre la valeur la plus récente en consultant la métadonnée (OGC 2008). D'autres approches réalisent des transformations sur les géométries des entités. Ware et Jones (1998) proposent une méthode de conflation qui calcule la géométrie d'un *feature* G du jeu de données résultant en faisant la moyenne des deux géométries appariées A et B. Cette moyenne dépend des sommets des géométries d'A et B et d'une valeur de tolérance d'erreurs combinée ε calculé pendant la phase d'appariement. (Ruiz et al. 2011) listent les techniques les plus utilisées dans la littérature. Par exemple : la technique de *rubber sheeting* qui utilise les triangulations de Delaunay (Gillman 1985) et la technique de Helmert (Watson 2006). La technique de *rubber sheeting* est l'application d'un champ vectoriel continu pour déformer l'ensemble des objets. Autrement dit, la géométrie de l'objet peut être considérée sous la forme d'une *membrane flexible* qui est ajoutée dans un cadre et obligée

donc à se déformer (Haunert 2005). Cet auteur propose une technique de *rubber sheeting* qui vise à préserver la topologie des réseaux des *features* linéaires (ex : routes, rivières, voies ferrées) et qui considère les liens d'ancrage entres des features représentant la même entité du monde réel avec différents niveaux de détail. Plus récemment, Touya et al. (2012) propose une amélioration de la technique de *rubber sheeting* en utilisant la méthode des moindres carrés afin de préserver les formes des objets. Certaines méthodes de *conflation* ont été implémentées dans des logiciels comme JCS Conflation Suite (JCS), 1Spatial Radius Studio, ACS$^{tm}$, ConfleX, ESEA MapMerger, GeoMedia Fusion (Ressler et al. 2009). Néanmoins, la plupart ont besoin de l'intervention de l'utilisateur afin de fournir des informations importantes pour la méthode de transformation.

# 4 Bilan de l'analyse des travaux existants

Cette section décrit les éléments, présentés dans cet état de l'art, qui nous semblent pertinents pour l'édition collaborative et pour la gestion de la cohérence d'un contenu géographique. Ces éléments sont résumés en Figure 33.

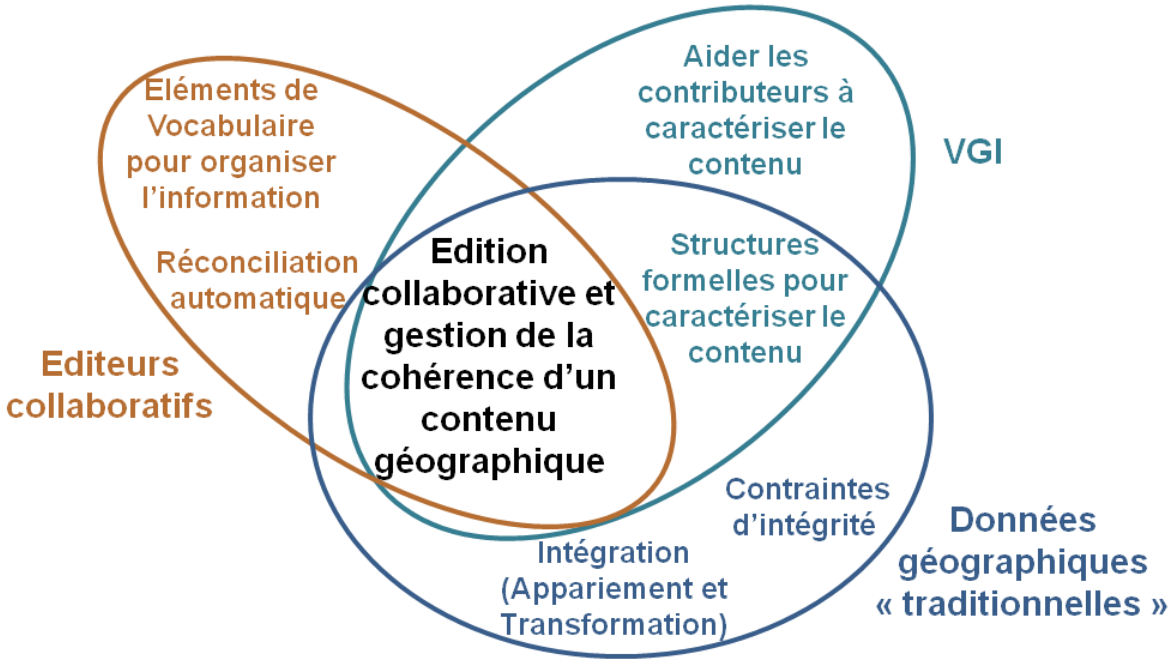



**Figure 33 : éléments dans les recherches décrites qui nous semblent pertinents pour l'édition collaborative et la gestion de la cohérence d'un contenu géographique**

Les éditeurs collaboratifs facilitent l'édition d'un contenu commun par un groupe de personnes. D'une part, les moteurs de wiki encouragent l'utilisation d'éléments de vocabulaire (dans le sens wiki) comme des catégories ou des modèles infobox, afin d'organiser l'information et permettre son traitement automatique. D'autre part, ces éléments permettent au contributeur d'éviter les incohérences de sens en utilisant d'homonymes. Ces éléments sont certaines fois prédéfinis dans le système ou dans la plupart de cas, ils sont définis par les contributeurs même. Les moteurs de wiki encouragent la réutilisation de ces éléments de vocabulaire partagés par les

utilisateurs, aidant à rendre le contenu homogène de manière à assurer une cohérence globale de ce contenu. Pour notre proposition, nous cherchons à identifier les éléments de vocabulaire pour un contenu géographique collaboratif. Nous nous intéressons également à permettre aux contributeurs de définir certains de ces éléments et d'encourager leur réutilisation. D'autre part, les éditeurs collaboratifs de modèles et de documents semi-structurés (XML) décomposent le contenu en considérant la notion de structure afin de pouvoir réconcilier des séquences d'éditions, concurrentes ou non, provenant d'utilisateurs différents. La réconciliation identifie les éditions indépendantes, celles sur des parties différentes du contenu, et conflictuelles, celles qui ne sont pas conformes à des contraintes. La réconciliation se sert souvent d'un historique d'éditions pour résoudre un conflit et pour pouvoir revenir vers des versions précédentes. Elle utilise aussi la notion des méthodes automatiques correctrices afin d'aider les utilisateurs à résoudre les conflits. Ces stratégies de réconciliation ont beaucoup traitées le cas de documents textuels, de modèles UML, et de graphes XML (Oster et al. 2007; Martin 2011; Michaux et al. 2011). Dans un premier temps, nous nous intéressons à la proposition de Michaux et al. (2011) sur les éditeurs collaboratifs de modèles UML car cette proposition considère la définition et l'évaluation de contraintes pour assurer la cohérence du contenu pendant la réconciliation.

L'utilisation des tags dans des contenus géographiques dits VGI, est une manière simple pour les contributeurs de caractériser les entités décrites. Ces tags s'approchent de l'idée d'un vocabulaire mais d'une nature non-structurée et non-formelle. L'utilisation des représentations formelles comme des ontologies ou des graphes RDF (Codescu et al. 2011; Ballatore et al. 2012) aident à mieux structurer/organiser ces tags. De cette manière, il est possible de permettre à un contributeur d'explorer ces structures formelles et de trouver des incohérences, comme par exemple des concepts redondants. Un autre aspect important pour la cohérence d'un VGI est l'utilisation de processus qui guident le contributeur pendant la caractérisation des entités, comme celui de suggestion des types de *features* proposé par Mülligann et al. (2011). Ces mécanismes semi-automatiques peuvent éviter l'introduction des incohérences dans le contenu en considérant les types d'objets et le contexte spatial des objets, nous savons par exemple qu'un bar dans une ville est probablement plus près d'une boite de nuit que d'une forêt. De plus, l'utilisation d'un jeu de données de référence pour l'évaluation de la qualité des données communautaires est utile pour effectuer des comparaisons automatiques des jeux de données et les améliorer. Le fait de permettre à un contributeur d'un projet communautaire de raccrocher son contenu à un jeu de données de référence devrait être supporté par ce projet.

La cohérence des données géographiques est traditionnellement gérée par la définition de spécifications pour leur acquisition et par l'évaluation des contraintes d'intégrité sur ces données. Ces contraintes d'intégrités sont fréquemment définies sous forme de relations spatiales qui doivent être préservées dans le contenu. Ces relations peuvent être dérivées des géométries d'objets par des opérations d'analyse spatiale (appelées des relations spatiales implicites) comme contient et disjoint. Dans un autre cas, ces relations peuvent être définies explicitement (appelées des relations spatiales explicites) dans le schéma de données de l'application qui est en charge de garder les liens explicites. Pour notre travail, nous retenons la classification de contraintes d'intégrité de Mäs et Reinhardt (2009) qui considèrent la typologie

suivante de relations spatiales : les relations topologiques (ex : contient, à l'intérieur, croise), les relations d'orientation (ex : à droite de, à gauche de, derrière, en face, au nord, au sud), les relations métriques (ex : près de, loin de). De plus, concernant l'évaluation de ces contraintes dans les données, nous retenons le travail de Servigne et al. (2000) qui fournissent des procédures de validation pendant l'acquisition des données afin de préserver la cohérence des données. Ces contraintes concernent le respect de la relation entre deux objets en prenant en compte leur nature dans le monde réel. Par exemple, un bâtiment est rarement à l'intérieur d'un autre bâtiment, mais un bâtiment peut être à l'intérieur d'une parcelle. D'autre part, la formalisation des spécifications de bases de données géographiques et l'adaptation des processus de *conflation* dans le cadre de l'édition collaborative de données géographiques peuvent aider à faciliter l'intégration des données et à gérer certaines incohérences liés à la mauvaise interprétation des spécifications par un néophyte.

Sur la base de ces éléments clés étudiés, nous construisons notre proposition décrite dans le chapitre suivant (chapitre III).

# Chapitre III

# *Coalla* : un modèle pour l'édition collaborative d'un contenu géographique et la gestion de sa cohérence

Dans ce chapitre, nous décrivons notre proposition de modèle baptisé *Coalla*[49], pour l'édition collaborative d'un contenu géographique avec gestion de la cohérence, en distinguant trois contributions.

La première contribution, présentée en section 1, est l'identification et la définition des éléments qui doit porter un vocabulaire formel partagé par des contributeurs, visant à faciliter la construction d'un contenu géographique collaboratif en aidant à assurer sa cohérence globale.

La deuxième contribution, présentée en section 2, est une méthode pour aider les contributeurs à construire certains de ces éléments en suggérant, à partir de mots-clés, des éléments de vocabulaire provenant de sources externes. Cette méthode s'appuie sur le contenu de Wikipédia France (et sa version structurée DBpedia) qui est une source collaborative de vocabulaire, une version formelle des spécifications IGN : une ontologie et un schéma produit de la base de données topographique BDTopo®, et la base de données linguistique WordNet (en français). Elle s'appuie également sur les tags OSM qui sont utilisés pour initialiser ce vocabulaire (détaillé dans la mise en œuvre).

La troisième contribution, présentée en section 3, est une stratégie d'évaluation et de réconciliation de contributions afin de les intégrer d'une façon cohérente à un contenu géographique collaboratif.

Enfin, nous présentons en section 4, une analyse comparative du modèle BDUni utilisé en production à l'IGN et notre modèle *Coalla*, après plusieurs discussions avec les experts du Service du Développement à l'IGN.

[49] Le modèle a été baptisé *Coalla* afin de souligner le mot « collaboration » tout en faisant référence au nom d'un animal, ici l'ours koala, qui se caractérise notamment par son système complexe de communication et d'organisation dans le but de préserver la cohésion sociale. Enfin, il s'agit aussi de faire référence aux livres d'O'Reilly, très réputés en informatique, qui sont également nommés par des noms d'animaux.

# 1 Éléments d'un vocabulaire formel pour la construction d'un contenu collaboratif cohérent

Cette section décrit notre modèle, illustré en Figure 34, d'un vocabulaire formel pour gérer la cohérence d'un contenu géographique collaboratif en appui sur un contenu de référence.

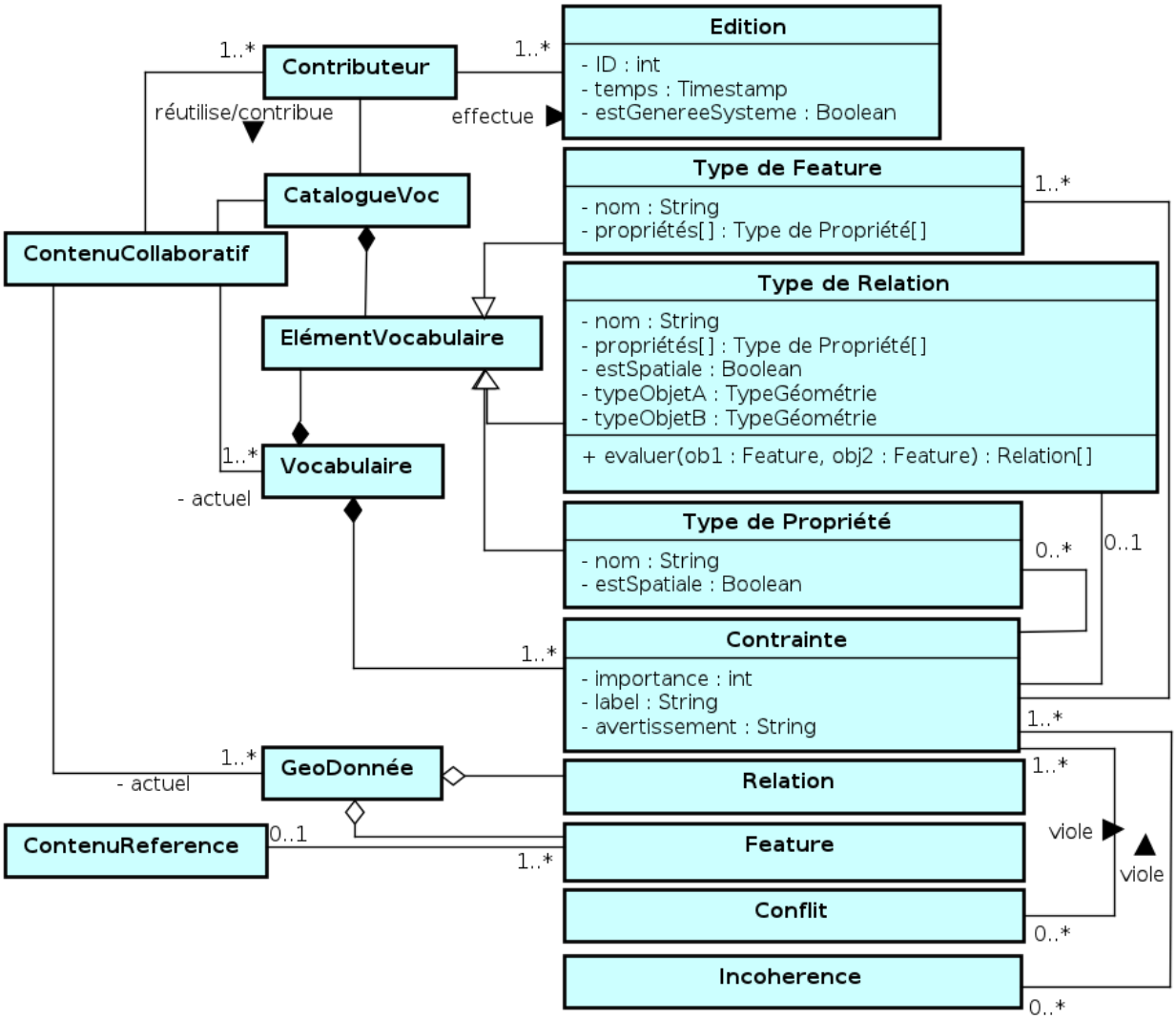



**Figure 34 : extrait du modèle conceptuel décrivant les éléments d'un vocabulaire formel pour la construction d'un contenu collaboratif cohérent**

Le `Contenu Collaboratif` est composé de plusieurs `GéoDonnées` (ex : un `GéoDonnées` sur le thème du tourisme, un `GéoDonnées` sur le thème des transports en commun). Un `Contributeur` travaille sur un ou plusieurs `GéoDonnées`. Les `GéoDonnées` sont des objets (`Feature`) et des `Relations`. Une `Relation` peut-être implicite et calculée par le système à partir des données. Une `Relation` peut aussi être explicitée entre un objet du contenu

collaboratif et un objet d'un Contenu Géographique de Référence. Cette relation sert à faire l'ancrage du contenu dans un référentiel de données IGN. En effet, l'IGN produit et diffuse des données géographiques de référence (RGE®) dont la qualité est connue et documentée. Ces données sont classiquement utilisées comme cadre de référence sur lequel d'autres données géographiques sont intégrées. Il semble pertinent d'étendre cet usage aux contenus collaboratifs.

Ce Contenu Collaboratif est caractérisé par son Vocabulaire (ex : un Vocabulaire sur le thème du tourisme, un Vocabulaire sur le thème des transports en commun). Un Contributeur travaille sur un ou plusieurs Vocabulaires. Un Vocabulaire est composé d'éléments d'un schéma de données géographiques, appelés Eléments de Vocabulaire. Plus précisément, un Type de Feature, un Type de Propriété, et un Type de Relation sont des méta-classes conformes à la norme ISO 19109 (ISO 2005) pour la description d'un schéma de données géographiques sous forme d'instances. En particulier, l'élément Type de Relation de notre modèle correspond à l'élément *AssociationType* défini sous cette norme ISO 19109 (ISO 2005). Un Vocabulaire est également composé de Contraintes exprimées au niveau du schéma que le contributeur veut voir respectées sur ce contenu. Une Contrainte peut être définie à partir d'un Type de relation entre un Type de Feature du Vocabulaire et un Type de Feature des spécifications de référence. Un Vocabulaire est construit en réutilisant des Eléments de Vocabulaire instanciés et répertoriés dans un Catalogue d'Eléments de vocabulaire. Ce catalogue est partagé par les contributeurs. Un Contributeur réutilise ces instances d'Eléments de Vocabulaire qui lui sont proposées selon ses besoins pour construire son schéma. Ces instances sont copiées dans son Vocabulaire si le contributeur le souhaite ainsi. Par exemple, le contributeur peut utiliser une instance de Type de Relation pour expliciter une Relation entre deux Features. Le Contributeur peut également créer un Elément de Vocabulaire. Cependant, le système l'encourage à réutiliser ceux dans le catalogue.

Un Contributeur effectue une Edition qui est une opération effectuée sur un Elément du Vocabulaire (ex : créer un Type de Feature dans le vocabulaire) ou sur une partie d'une donnée (ex : modifier la propriété d'un Feature ou d'une Relation).

## 1.1 Les types de *features*

L'organisation des entités géographiques du monde réel en types de *features* est une façon classique et intuitive de modéliser l'espace (Raper 1996; Parent et al. 1998; Mennis et al. 2000; Bédard et al. 2004; ISO 2005). Pour cette raison, nous avons choisi de garder ce principe dans la proposition. Nous proposons de garder la classe générique Thing afin de donner la liberté au contributeur d'instancier un objet en saisissant sa forme ou en signalant au moins sa présence sans en connaitre sa nature.

## 1.2 Les propriétés et relations spatiales potentiellement pertinentes

Les relations spatiales sont une composante clé de l'espace géographique. En effet, elles jouent un rôle important dans la manière dont les gens perçoivent, raisonnent, et décrivent l'information spatiale (Egenhofer et Mark 1995). De même, certaines propriétés spatiales d'objets comme la forme et la taille sont importantes dans notre perception et discrimination des objets remarquables (Deakin 1996). Certaines relations topologiques, d'orientation et certaines propriétés désignant un objet remarquable permettent l'interprétation automatique des observations dessinées par les gens sur leur espace dans un croquis cartographique (Chipofya et al. 2011; J. Wang et al. 2011). Certains de ces relations et propriétés jouent un rôle important dans des modèles de ville 3D pour les applications SIG urbaines, par exemple, la rue qui est entourée de très hauts bâtiments, un aéroport qui est proche d'une ville, ou encore la hauteur et la largeur d'une porte (Bucher et al. 2012). Les relations et les propriétés spatiales sont de fait utilisées pour la gestion de l'intégrité de bases de données géographiques (Hadzilacos et Tryfona 1992; Borges et al. 2002) et sont un critère important dans les processus de changement de niveau de détail ou de cartographie en tant qu'information à préserver (Ruas et Plazanet 1996; Touya et al. 2010). Globalement, l'explicitation des propriétés et des relations spatiales potentiellement pertinentes pourrait rendre le contenu plus utile, dans le sens d'usage des données selon la norme ISO 19113 (ISO 2002), c'est-à-dire la qualité externe, ainsi que pour aider à la gestion de la cohérence du contenu.

Le rôle des propriétés et relations dans l'édition collaborative peut être d'aider à détecter les incohérences dans les données et possiblement à les résoudre. Le fait de les connaître explicitement (relations explicites) ou des les calculer grâce aux géométries (relations implicites) nous permet de vérifier automatiquement qu'elles sont toujours préservées dans le contenu. La Figure 35 illustre ces notions en comparant deux représentations possibles (R1 et R2) d'un même bâtiment du monde réel (symbole de couleur noire). R1 est une représentation cohérente vis-à-vis de la forme, de l'orientation et de la distance au virage. R2 est une représentation cohérente en regard de la relation topologique avec la route. Selon l'application, les relations et propriétés importantes à préserver ne seront pas les mêmes. Dans le cas général, la relation topologique entre un bâtiment et le réseau routier prime et c'est la représentation R2 qui sera choisie, comme celle présentant des incohérences non graves, alors que R1 sera éliminée. Ces relations et propriétés sont rarement décrites explicitement mais peuvent être dérivées à partir des coordonnées (géométries) des objets. Dans l'exemple de la Figure 35, les objets sont décrits par des attributs thématiques et par une géométrie (polygone pour les bâtiments et ligne pour la route). Les relations et propriétés intervenant dans l'évaluation de la cohérence sont calculées à partir des géométries.

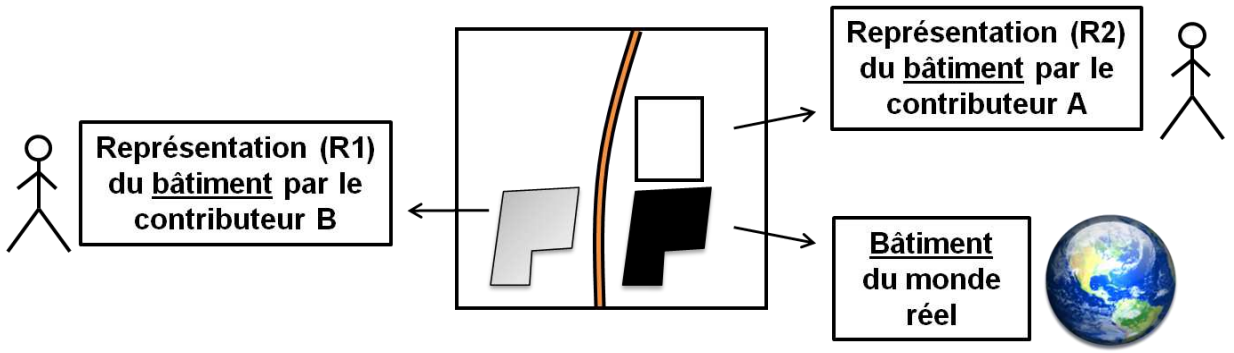



**Figure 35 : deux représentations R1 et R2 du bâtiment du monde réel (en noir) présentant des incohérences vis-à-vis de la forme ou des relations spatiales, plus ou moins graves selon l'application**

Les gens utilisent souvent certaines relations spatiales et propriétés d'objets remarquables afin de se repérer dans un espace géographique (Deakin 1996; Jaara et al. 2012), par exemple, « l'église est en face du bureau de poste » ou « le magasin de vêtement est sur la place ». Un contributeur peut avoir le besoin de signaler des relations et propriétés spatiales potentiellement pertinentes comme par exemple « la rivière qui longe la forêt » ou « l'unique maison de façade bleue au milieu des maisons de façades blanches ». Nous avons émis l'hypothèse que, pour certains contributeurs, il est parfois mieux de s'exprimer en regard des relations et des propriétés (informations qualitatives) que des géométries (information quantitative) afin de décrire une contribution. Aussi, une hypothèse plausible est que cette intuition est toujours valide dans le contexte de l'édition collaborative. Nous avons posé la question dans un sondage proposé à des contributeurs d'OSM sur leurs avis concernant le renseignement de certaines relations[50] comme « l'abribus est exactement en face du bureau de poste », ou de certaines propriétés comme « c'est le plus grand immeuble de la zone ». Il s'est avéré que les contributeurs d'OSM ne souhaitaient pas s'exprimer ainsi. Sur ce point, un contributeur explique « *la position relative des objets est fournie par la position des points sur la carte, et les objets peuvent être décrits par des attributs individuels (hauteur, nombre d'étage)* ». Ils étaient principalement intéressés d'expliciter des relations spatiales pour faciliter la mise à jour en les utilisant comme des alertes, par exemple le cas où un nouvel immeuble est construit ou que un abribus est déplacé". Elles sont considérées utiles pour les zones où il manque des points de référence et notamment pour faciliter l'orientation et le repérage dans une application d'aide à la navigation. Ils considèrent que ce type d'information est subjectif et pourrait donc être utile dans le cadre d'une application grand public. Sur ce point, un contributeur indique :

« *J'aimerais pouvoir déterminer quels sont les 2 ou 3 principaux repères géographiques à proximité d'un lieu donné. Ceci est nécessaire pour produire des plans d'accès à des bâtiments publics. En effet, il faut d'abord localiser le bâtiment cible puis, pour choisir un périmètre de plan qui soit utile aux visiteurs, il faut choisir un périmètre qui soit suffisamment grand pour inclure un repère géographique très connu à proximité. Par exemple, pour faire le plan d'accès à la mairie*

[50] Nous avons clarifié que notre notion de relation est différente à celle dans OSM (voir section I.3.2).

*de Boulogne-Billancourt, il est souhaitable d'y inclure le MacDonald's le plus proche et la station de métro la plus proche. Le problème, c'est que la notion de repère principale est subjective et est généralement connue des habitants. En leur demandant : pourriez-vous m'expliquer comment me rendre dans ce lieu, ils utilisent des repères qu'ils estiment connus ou facilement repérables (Mac Do, métro, etc.)* ».

En effet, ces relations et propriétés sont sans doute précieuses pour améliorer l'utilisabilité des contenus pour les applications de, par exemple, aide à la navigation. Elles doivent être préservées afin de garantir la cohérence d'un contenu. Cependant, il est important d'acquérir ces relations et propriétés autrement que via le contributeur.

## 1.2 Les types de propriétés et les types de relations

Nous avons prédéfini dans le *catalogue d'éléments de vocabulaire* des types de relations (binaires asymétriques) et des types de propriétés. Nous les avons associés des méthodes pour les évaluer automatiquement dans les données. Lorsque le contributeur construit son vocabulaire, il peut reprendre ces types de relations ou de propriété prédéfinis. De cette manière, il est également possible d'encourager la réutilisation des éléments du vocabulaire dans le catalogue.

Un `Type de Relation` correspond à l'élément *AssociationType* défini sous la norme ISO 19109 (ISO 2005). Un `Type de relation` peut être défini entre deux classes, l'une correspondant à des objets surfaciques et l'autre à des objets linaires. Il existe par exemple, une spécialisation de la classe `Type de relation` appelé `intersecte` entre des objets surfaciques et des objets linéaires. La Figure 36 montre le type prédéfini de relation `intersecte`.

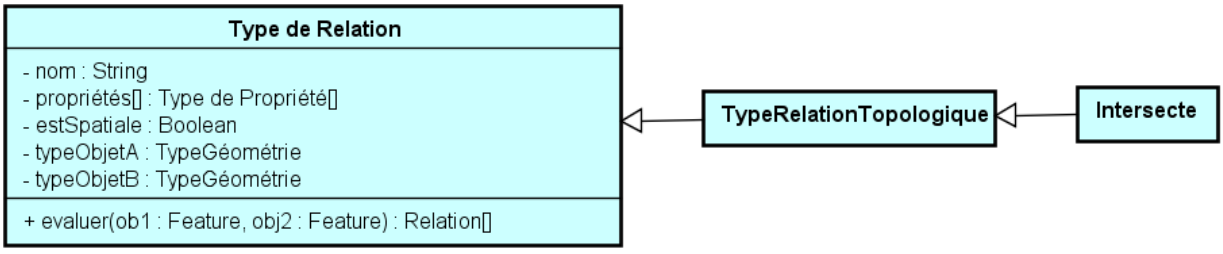



**Figure 36 : un type prédéfini de relation `intersecte`**

Un `Type de propriété` peut correspondre à une ou plusieurs classes, par exemple, le type de propriété `espèce d'arbres` pour des objets appartenant à des classes : réserve naturelle et jardin botanique. Un contributeur peut définir un type de propriété dont la valeur sera entrée à la main, par exemple la propriété `nom` pour des entités nommées comme les fontaines ou les églises. Un contributeur peut également définir un type de propriété dont la valeur sera calculée

à partir d'une géométrie, par exemple, la longueur, le centre, et la surface d'une forme polygonale, la longueur d'un segment, etc. Il existe une spécialisation de la classe `Type de propriété` qui correspond à la longueur d'un type d'objet linéaire. La Figure 37 montre le type prédéfini de propriété `longueur`.

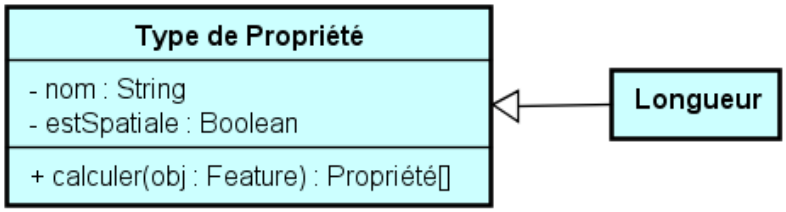



**Figure 37 : un type prédéfini de propriété `longueur`**

La propriété booléenne `est spatiale` d'un `Type de propriété` et d'un `Type de relation` est vraie lorsque les instances de ce type de propriété et de relation peuvent être déterminées à partir de la géométrie des objets. Le type de propriété possède alors une méthode pour la calculer. De la même manière, le type de relation comporte alors une méthode spécifique pour l'évaluer. Ces méthodes dépendent du type de géométrie des classes concernées et ont été implémentées sur la plate-forme GéOxygène du Laboratoire COGIT (Grosso et al. 2012).

## 1.3 Les contraintes sur des types de relations

Dans notre modèle, une `Contrainte` peut être définie sur des types de *features* (au niveau du modèle). Elle peut représenter une vérité indiscutable ou vérifiée le plus souvent sur la configuration de notre espace, par exemple, « un bâtiment ne flotte pas sur l'eau » ou « un bâtiment ne chevauche généralement pas une route ». Dans une zone urbaine, ce sont les règles d'urbanisme définies par les mairies, par exemple la disposition des bâtiments par rapport à leur rue est réglementée dans le Plan local d'urbanisme de la ville de Paris[51]. Une `Contrainte` peut également être spécifique sur des objets géographiques (au niveau des instances), par exemple, « cette maison est du même coté de la rue que l'abribus ».

Une `Contrainte` s'exprime à l'aide d'un type de relation. Elle possède une étiquette afin de la nommer et de la stocker dans un *catalogue de contraintes* (détaillé dans la mise en œuvre) disponibles dans le système. Elle a aussi un message prédéfini d'avertissement expliquant en langage naturel l'incohérence que cette contrainte détecte, de même qu'un niveau d'importance (une valeur qui est optionnelle) indiquant la priorité d'évaluation de la contrainte. Une contrainte peut représenter une règle d'interdiction comme `n'intersecte pas` et peut également avoir

[51] http://www.paris.fr/pratique/urbanisme/documents-d-urbanisme-plu/p6576

des paramètres comme la distance pour la contrainte `est à une distance supérieure à`. Nous avons d'ailleurs prédéfini dans le système ces deux contraintes.

La classe `Contrainte` joue un rôle central dans la gestion de la cohérence du contenu. La violation d'une contrainte est une `Incohérence` qui est caractérisé par la contrainte elle-même et les objets qui ne la satisfont pas. L'`Incohérence de Relation Spatiale` est calculé grâce à un catalogue de méthodes définies dans une classe utilitaire (détaillé dans la mise en œuvre). Pour certaines contraintes, des méthodes de correction d'incohérences ont été implémentées sur GéOxygène (Grosso et al. 2012) et utilisent des méthodes d'analyse spatiale disponibles sur la plate-forme de généralisation cartographique CartAGen du Laboratoire COGIT (Renard et al. 2011). En effet, notre modèle s'est inspiré de modèles de généralisation cartographique qui visent également à préserver des relations importantes (Ruas et Mackaness 1997; Duchêne 2004). Notre modèle ne prend pas en compte des situations complexes comme celle gérées très fréquemment en généralisation cartographique, par exemple, des contraintes de densité du réseau de rues sur une ville (*contrainte composite ou meso*), des groupes de maison trop proches ou des lacs trop petits (Cécile Duchêne 2004). Il est possible de réutiliser ces méthodes d'analyse spatiale disponibles dans CartAGen pour implémenter les méthodes de correction. La Figure 38 montre la représentation d'une contrainte et la contrainte prédéfinie `n'intersecte pas`.

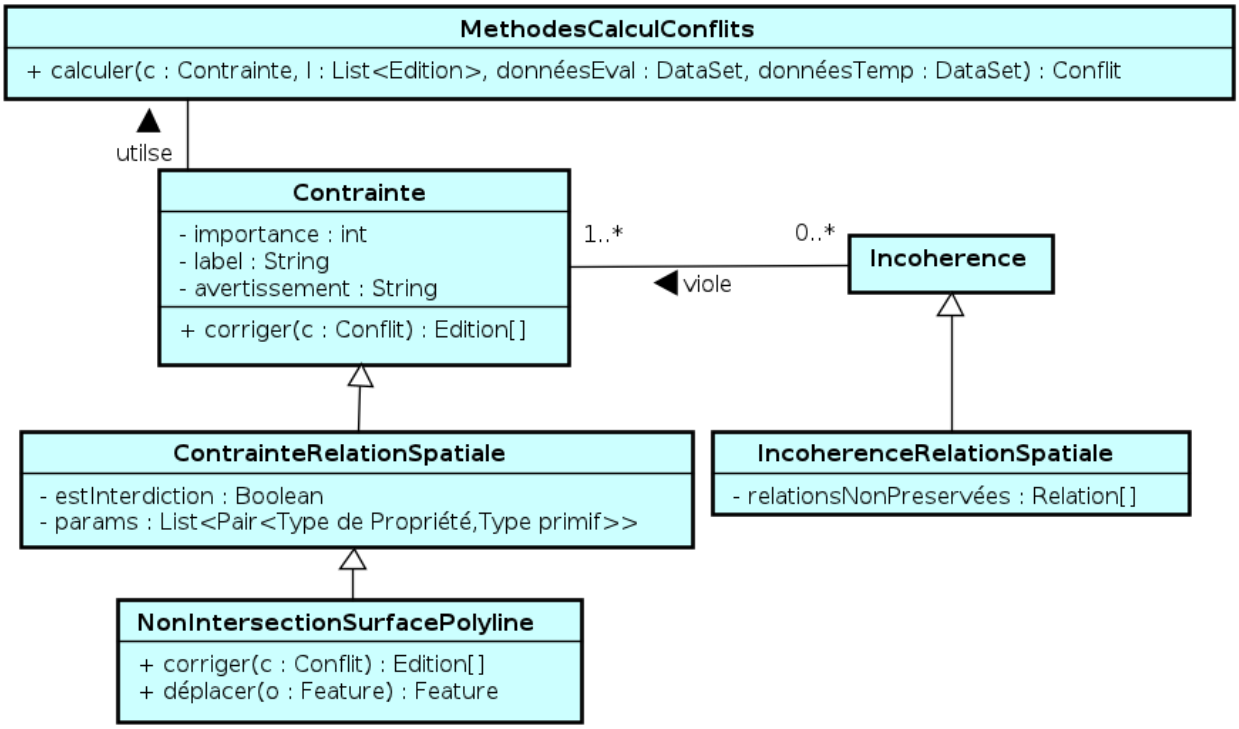



**Figure 38 : la contrainte prédéfinie `n'intersecte pas` et son incohérence correspondante dans le cas de violation de cette contrainte**

L'intérêt principal de ces contraintes est de pouvoir définir des types de relations importants entre une classe du schéma contributeur et une classe BDTopo® ou un concept IGN. Il existe deux intérêts principaux derrière cette idée. D'une part, il est souhaitable de montrer au contributeur la façon dont une agence gouvernementale de cartographie construit son schéma de données. D'autre part, il est également souhaitable que le contributeur construise des références vers des classes IGN prédéfinies qui pourrait servir pour définir des contraintes d'intégrité entre une classe IGN et une classe collaborative. Par exemple, le contributeur via l'interface d'édition pourrait établir une contrainte sur le type de relation `est dans` entre les objets d'une classe du contenu collaboratif comme `Restaurant` et les objets d'une classe IGN comme `Bâtiment`. Le respect de cette relation, c'est-à-dire que les restaurants sont situés dans des bâtiments pourrait être systématiquement vérifié par le système.

## 1.4 Les contraintes de dépendance entre des types de propriétés

Sauf si cela est précisé autrement, les propriétés d'un objet sont toutes indépendantes les unes des autres, par exemple le nom et la géométrie d'un tronçon de route peuvent être édités indépendamment. Afin d'assurer la cohérence du contenu pendant l'édition concurrente du contenu, il est également possible de préciser la dépendance entre plusieurs types de propriétés, par exemple la longueur d'un tronçon de route dépend de sa géométrie. De cette façon, l'édition concurrente des propriétés indépendantes ne produit pas un conflit. Dans le cas contraire, le système détecte un conflit. La Figure 39 détaille la représentation d'une contrainte de dépendance dans le modèle. Un contributeur pourrait s'il le souhaite, définir une contrainte de dépendance entre le type de propriété `longueur` et `géométrie`.

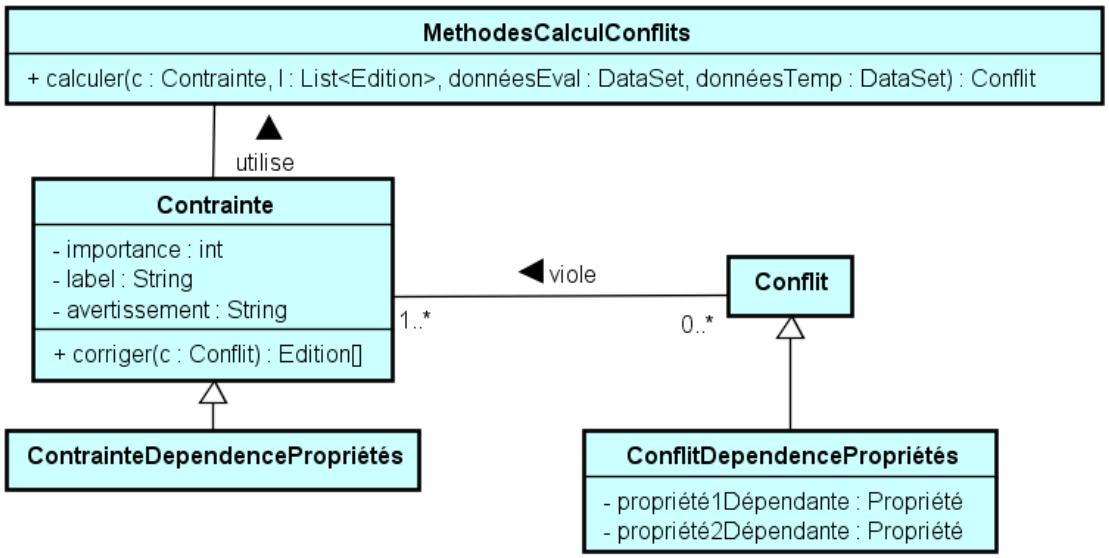



**Figure 39 : la contrainte prédéfinie de dépendance entre des types de propriétés longueur et géométrie et le conflit correspondant dans le cas de violation de cette contrainte**

# 2 Une méthode pour aider les contributeurs à la construction du vocabulaire formel[52]

Cette section présente un processus semi-automatique d'aide à la construction de certains éléments du vocabulaire selon le besoin d'un contributeur. La motivation de concevoir ce processus est de fournir les contributeurs avec des éléments de vocabulaire prédéfinis et possiblement pertinents à leurs besoins. Certains de ces éléments ont été manuellement prédéfinis dans le *catalogue d'éléments de vocabulaire*, comme le type de relation `intersecte`. Cependant, il existe un clair besoin d'acquérir ces éléments automatiquement, et pour cette raison nous avons conçu ce processus. Le contributeur déclenche le processus afin de mettre à jour son vocabulaire, un processus différent que celui de mise à jour de ses données. Autrement dit, le contributeur construit le modèle et ensuite met à jour ses données en s'adhérant à ce vocabulaire.

La Figure 40 illustre globalement la manière dont un contributeur interagit avec notre processus.

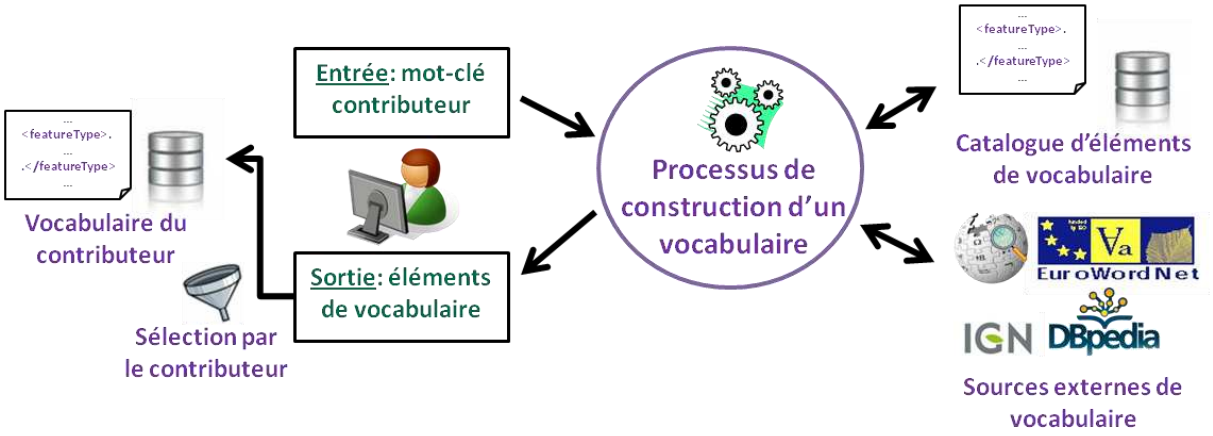



**Figure 40 : processus semi-automatique pour la construction d'un vocabulaire formel (général)**

Le contributeur exprime son besoin par des mots clés décrivant la nature ou la fonction de la donnée souhaitée. Un mot-clé peut être simple (ex : autoroute) ou composé (ex : ligne de métro). En réponse à ces mots-clés, notre processus propose des classes et des types de propriété et relation, ainsi que des éléments du RGE auxquels raccrocher son contenu. Ces éléments sont de préférence extraits à partir du *catalogue d'éléments de vocabulaire*. En absence de réponse, le processus se sert des *sources externes de vocabulaire* : des sources collaboratives comme Wikipédia (France) et sa version structurée DBpedia et des sources expertes comme des spécifications formelles IGN et la base de données linguistique WordNet

[52] Certaines parties de cette section sont adaptées de :
Brando C, Bucher B, Abadie N, Specifications for User Generated Spatial Content, in: Geertman S, Reinhardt W, Toppen F (eds) Advancing Geoinformation Science for a Changing World, Springer-Verlag Lecture Notes in Geoinformation and Cartography, pp 479-495, Utrecht, Netherlands

(en français). Pour peupler notre catalogue, la plupart des éléments ont été extraits par ce processus à des *sources externes de vocabulaire* en lui donnant en entrée des tags OSM. Le contributeur choisit ensuite, via une interface graphique, les types de *features*, types de propriétés, et types de relations, pertinents selon son besoin. Les éléments choisis sont ensuite conservés dans le vocabulaire de même que dans le *catalogue d'éléments de vocabulaire* pour qu'ils puissent être réutilisés par d'autres contributeurs.

Nous détaillons ensuite la partie de processus qui extrait des éléments de vocabulaire à partir des *sources externes de vocabulaire*. Plus précisément, le processus extrait des éléments Wikipédia, éléments DBpedia, éléments WordNet, et des éléments IGN. Les différentes parties du processus et leurs enchaînements sont illustrés en Figure 41.

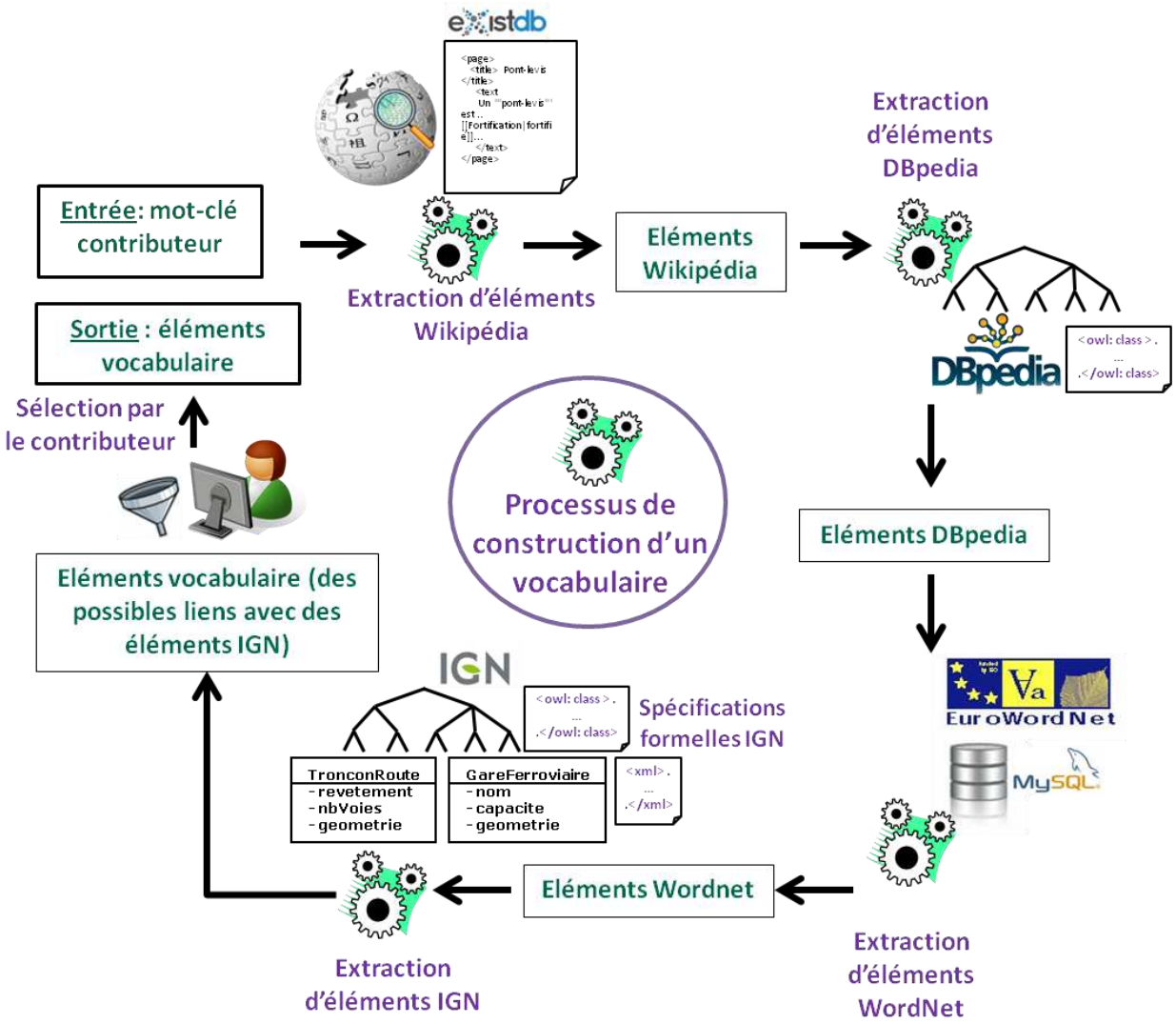



**Figure 41 : processus semi-automatique pour la construction d'un vocabulaire formel (à partir des vocabulaires externes)**

La Figure 42 présente les correspondances entre les éléments extraits par le processus et le vocabulaire à construire. Les détails des correspondances sont expliqués plus après.

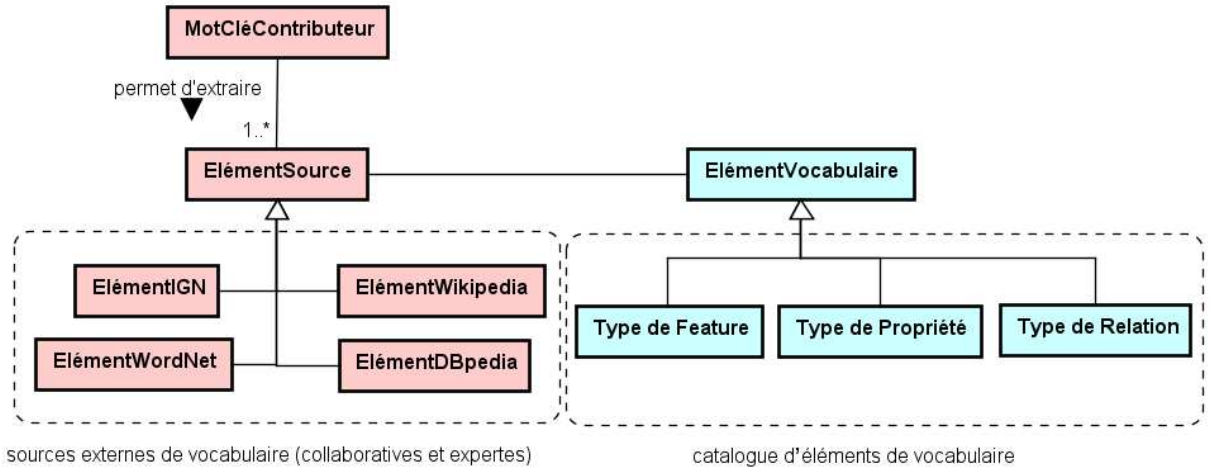



**Figure 42 : les correspondances entre les éléments extraits des sources externes et le vocabulaire à construire**

Pour illustrer les étapes du processus dans la suite, nous prendrons des exemples de résultats produits par la mise en œuvre à décrire dans le chapitre 4, à partir du mot-clé `aéroport`.

## 2.1 Extraction d'éléments Wikipédia : types de *features* et types de propriétés pour le vocabulaire

Wikipédia semble être une source importante pour un vocabulaire géographique. En effet, 526.000 lieux existent dans l'encyclopédie[53] et sont décrits, dans l'article correspondant, en utilisant des catégories inclues dans le graphe de catégories de Wikipédia (WCG) et des types de propriétés définies dans les modèles *infobox*. Il s'avère qu'une partie importante des contributeurs d'OSM sont aussi des contributeurs actifs à Wikipédia (Budhathoki, 2010). Il est peut être donc possible de trouver un vocabulaire en commun entre les deux projets communautaires. A partir du mot-clé du contributeur, le processus extrait des *éléments Wikipédia*, en premier lieu, des catégories du WCG et ensuite, des types de propriétés à partir des modèles infobox. La Figure 43 illustre les *éléments Wikipédia* extraits et leurs correspondances avec les *éléments de vocabulaire* à construire.

[53] La version 3.7 de BDpedia contient 526.000 lieux extraits à partir Wikipédia entre les 1.83 millions d'entités : http://blog.dbpedia.org/2012/08/06/dbpedia-38-released-including-enlarged-ontology-and-additional-localized-versions/

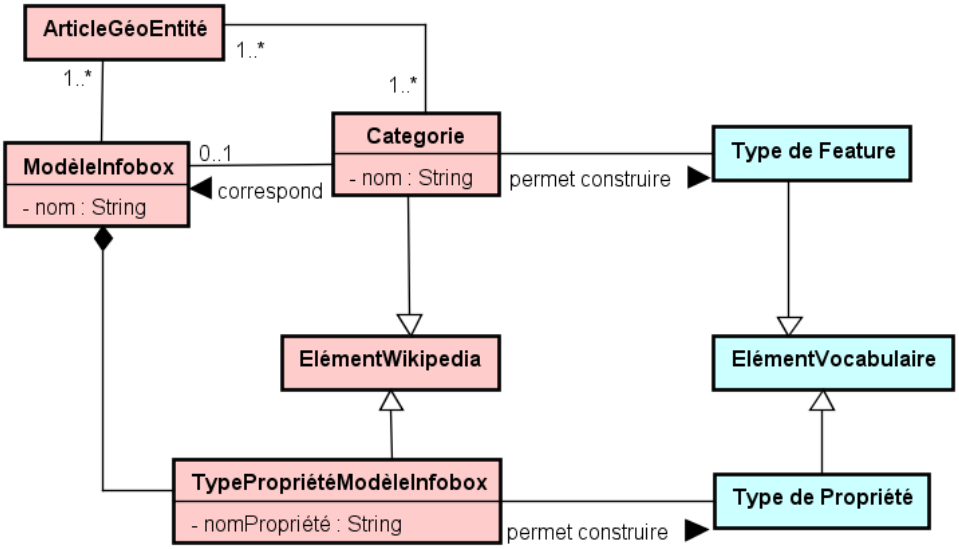



**Figure 43 : les éléments du vocabulaire qui sont construits à partir des éléments Wikipédia**

### 2.1.1 Extraction de catégories

Afin d'extraire les catégories, l'extracteur interroge le WCG afin de trouver celles qui ressemblent syntaxiquement au mot-clé, plus précisément, les catégories dont le nom contient le mot-clé. Un contributeur pourrait vouloir spécifier un simple mot-clé comme `montagne` ou `équipement public` pour obtenir des catégories générales ou non comme `refuge de montagne des Alpes`, pour obtenir des catégories spécifiques. Une fois une catégorie trouvée, l'extracteur a deux choix possibles. Dans un premier cas, l'extracteur cherche les catégories sans inclure les sous-catégories dans le résultat. Dans l'autre cas, l'extracteur cherche les catégories et inclut les sous-catégories correspondantes (à partir du WCG) dans le résultat. Par exemple, la catégorie générale `Cours d'eau` contient des sous-catégories comme `Delta`, `Rivière Souterraine`, `Canal`, `Fleuve`, et `Créature des cours d'eau` (voir l'extrait du WCG sur la Figure 44). Le contributeur devra choisir la catégorie et les sous-catégories qui sont pertinentes pour son contenu.

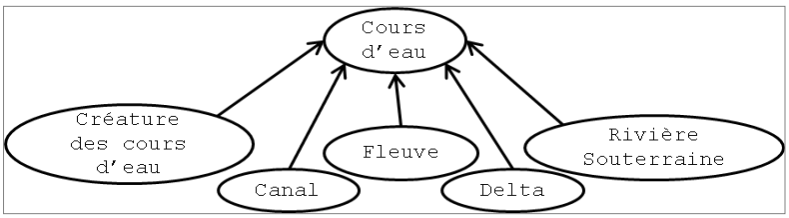



**Figure 44 : extrait du WCG correspondant à la catégorie `Cours d'eau` et ses sous-catégories**

La catégorie sélectionnée correspond à un type de *feature* (voir la Figure 43) dont son nom est initialisé avec le nom de la catégorie. De la même manière, un type de *feature* par sous-catégorie est créé. Un type de relation `est un` est défini entre le type de *feature* correspondant à la catégorie, et les types de *features* correspondant aux sous-catégories. De plus, chacune contient l'information détaillée sur la source afin de permettre au contributeur de connaître l'origine de cet élément ainsi que la façon dont il a été retrouvé. Cette information d'origine correspond au mot-clé utilisé dans la recherche, au titre de la page (nom de la catégorie) et aux catégories mères correspondantes. Si la catégorie `aéroport` est sélectionnée par le contributeur, le type de feature est créé. La Figure 45 illustre les éléments de vocabulaire choisis pour l'instant par notre contributeur. Ces éléments seront utilisés aussi pour la prochaine phrase du processus.

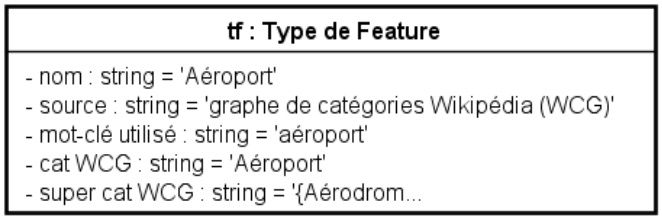



**Figure 45 : type de feature construit à partir du mot aéroport (résultat issu de la mise en œuvre décrite en chapitre 4)**

### 2.1.2 Extraction des types de propriétés

L'étape suivante est de trouver des types de propriétés pour le type de feature identifié. L'extracteur se sert des noms de catégories et de sous-catégories précédemment sélectionnées afin d'obtenir les modèles infobox dont leurs noms sont ressemblant syntaxiquement. Plus précisément, les modèles infobox dont les noms contiennent le nom de la classe identifiée. Par exemple, le modèle infobox appelé `Modèle:Infobox Voie Parisienne` est trouvé grâce à la sous-catégorie du même nom `Catégorie:Voie Parisienne` appartenant à la catégorie `Catégorie:transport à Paris`. Ensuite, les contenus des modèles infobox trouvés sont traités afin d'extraire leurs champs. La Figure 46 montre par exemple un extrait de la définition en wiki-code du `Modèle:Infobox Refuge` avec quelques champs intéressants comme `nom`, `altitude`, `région`, `classement`, `capacité été`, `gérant` et l'utilisation de cette infobox dans l'article `Refuge de Pombie`.

```
{{Infobox Refuge
  |nom=
  |altitude=
  |massif=
  |pays=
  |région=
  |subdivision=
  |propriétaire=
  |gérant=
  |période=
  |capacité été=
  |capacité hiver=
  |classement =
  |latitude=
  |longitude=
  …
}}
```


**Figure 46 : (à g.) l'extrait de la définition en wiki-code du `Modèle:Infobox Refuge` et (à d.) extrait de cette infobox `Refuge` utilisée dans l'article `Refuge de Pombie`**

A partir de chaque champ du modèle infobox, un type de propriété est créé pour le type de feature précédemment créé. Un type de propriété possède un nom et contient aussi l'information détaillée sur la source afin de permettre au contributeur de connaître l'origine de cet élément et comment il a été retrouvé. Cette information d'origine correspond au mot-clé utilisé dans la recherche, au nom du modèle infobox et au nom du champ correspondant. A partir de l'exemple sur les aéroports, le contributeur devra choisir les champs de l'infobox qui sont pertinents pour son besoin. Disons qu'il choisit les champs `nom`, `gestionnaire`, et `superficie` provenant du modèle `Infobox:Aéroport`. Donc, trois types de propriétés sont initialisés et associés au type de feature créé dans l'étape précédente. La Figure 47 illustre les éléments de vocabulaire choisis pour l'instant par notre contributeur.

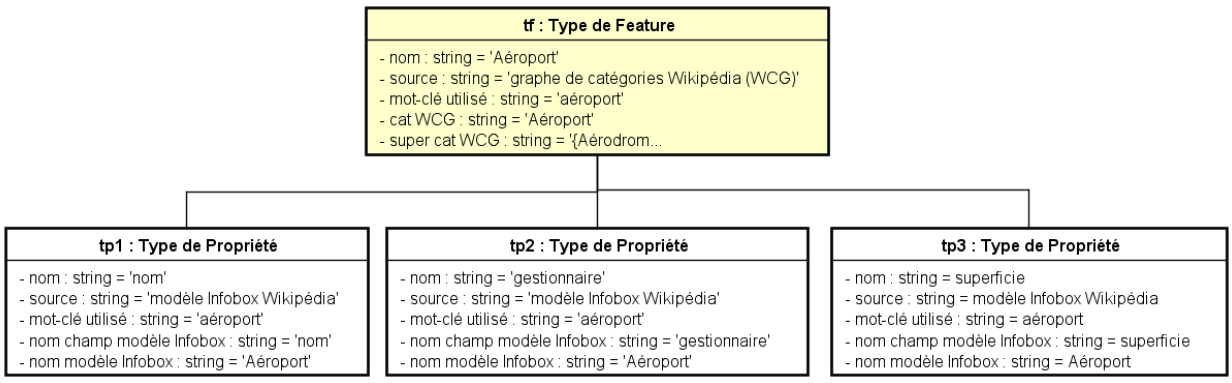



**Figure 47 : trois types de propriétés du vocabulaire construit à partir du mot aéroport (résultat issu de la mise en œuvre décrite en chapitre 4)**

## 2.2 Extraction d'éléments DBpedia : types de propriétés et de relation pour le vocabulaire

Un modèle formel décrivant une partie importante de contenu Wikipédia existe sous la forme de l'ontologie DBpedia (Bizer et al. 2009). Elle est une source importante de types de relations et propriétés existantes dans Wikipédia. A partir du mot-clé et des catégories (et sous-catégories) trouvées dans la première étape du processus, il est possible d'extraire des éléments DBpedia pour le vocabulaire, précisément des types de propriétés et des types de relations potentiellement pertinents. La Figure 48 illustre les *éléments DBpedia* extraits et leurs correspondances avec les *éléments de vocabulaire* à construire.

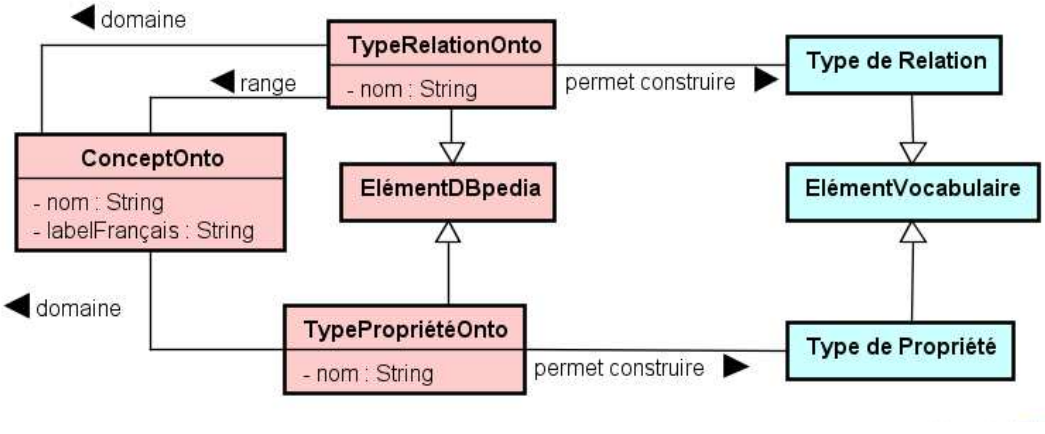



**Figure 48 : les éléments DBpedia extraits et leurs correspondances avec les éléments de vocabulaire à construire**

### 2.2.1 Description de l'ontologie DBpedia

L'ontologie DBpedia anglaise a été manuellement dérivée à partir des infoboxes de la version anglophone de Wikipédia. Elle est enrichie au fur et à mesure par les administrateurs du projet DBpedia et des nouvelles versions sont sorties périodiquement. Les concepts de haut niveau de DBpedia sont au nombre de 5 : `Person`, `Place`, `Organization`, `Specie`, et `Disease`. Au dessous du concept `Place`, il existe onze concepts directs, dix types de propriétés et types de relations (entre `DatatypeProperty` et `ObjectProperty`). La Figure 49 montre un extrait de cette ontologie.

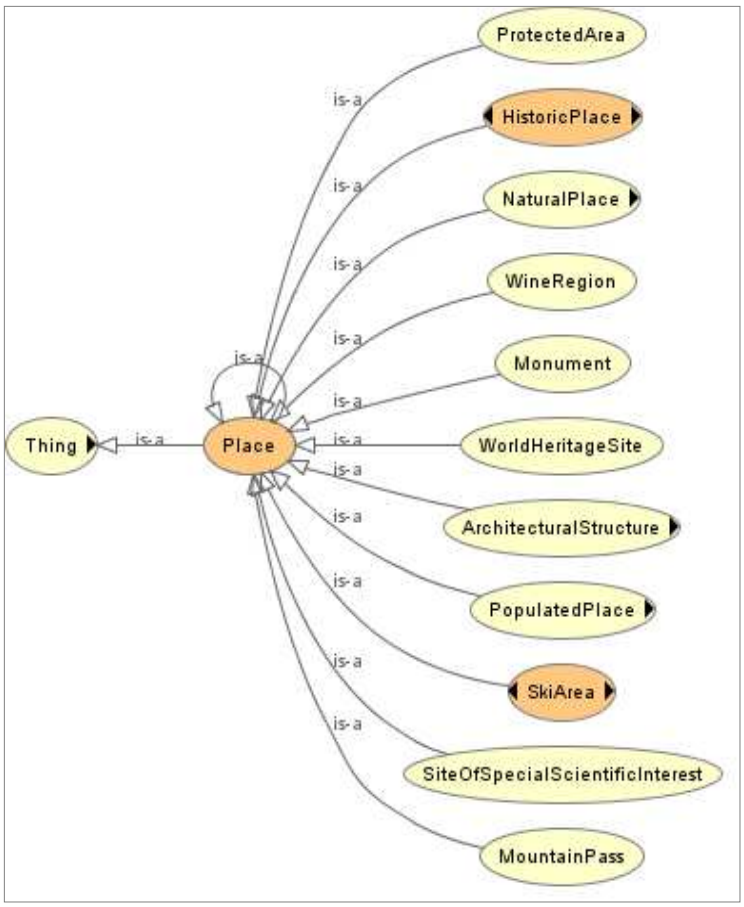



**Figure 49 : un extrait du graphe DBpedia 3.7 concernant le concept Lieu (Place en anglais)**

L'utilisation d'une ressource formelle comme l'ontologie DBpedia permet la distinction entre des types de propriétés et des types de relations des modèles infobox. Globalement, la plupart des champs de l'infobox peuvent être considérés comme des types de propriétés. Néanmoins, il est possible de trouver plus récemment des champs qui correspondent à des types de relations, par exemple, le type de relation `capital city` du `Modèle:Infobox Country` peut avoir comme valeur l'URI dans Wikipédia. Cette URI présente l'intérêt d'être déréférençable et de conduire vers un article. Par exemple, l'infobox `Country` est utilisée dans l'article « France » où le champ `capital city` conduit à l'article « Paris ».

En considérant que notre processus est adapté pour utiliser uniquement des ressources en français, nous avons considéré le très récent projet DBpedia en français[54]. Il fournit des données brutes sous la forme de triplets (*dumps*) qui sont extraites à partir du contenu de la version francophone de Wikipédia. Il faut clarifier qu'il n'existe pas encore une ontologie DBpedia française adaptée au contenu de la version française de Wikipédia. Il est néanmoins possible de trouver des labels en français `label@fr` (et dans d'autres langues occidentales) pour environ plus de la moitié de concepts de DBpedia anglophone. Nous avons manuellement ajouté des labels en français via l'éditeur d'ontologies OWL Protégé[55] pour le reste des

[54] http://www.dbpedia.fr/
[55] http://protege.stanford.edu/

concepts. Nous avons traduit les concepts de l'anglais au français grâce aux liens multilingues présents fréquemment dans les pages Wikipédia.

En considérant que DBpedia est principalement dérivée des infoboxes de Wikipédia, nous avons décidé de garder les modèles infobox (cf. sous-section III.2.1.3) pour la construction de notre vocabulaire car ils étaient en français et certaines infoboxes n'ont pas encore été incorporées dans l'ontologie DBpedia anglaise. De plus, il existait la possibilité de choisir les concepts de l'ontologie comme des classes potentielles de notre vocabulaire. Néanmoins, nous avons décidé de garder le WCG de la version française de l'encyclopédie car il est adapté spécifiquement au contenu de la version francophone de Wikipédia et est très exhaustif vis-à-vis du nombre de classes potentielles.

### 2.2.2 Extraction des types de propriétés et des types de relations

A partir du mot-clé et des catégories (et des sous-catégories, si elles sont indiquées) trouvées dans la première étape du processus, il est possible d'extraire des types de propriétés et des types de relations associés à des concepts DBpedia. Plus précisément, les types de propriétés et les types de relations des classes DBpedia dont les labels sont syntaxiquement similaires au mot-clé, aux catégories ou aux sous-catégories. Pour élargir les résultats de la recherche, nous considérons également le concept mère du concept DBpedia trouvé. La comparaison entre chaînes de caractères se fait grâce à la mesure de similarité N-Gram (N=3). Nous avons testé la distance lexicale de Levenshtein qui est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre (Euzenat and Shvaiko, 2007). La mesure lexicale par N-Gram était empiriquement la plus satisfaisante car elle est plus efficace pour comparer une chaîne de caractères contenant fréquemment une autre chaîne de caractère. Par exemple, la comparaison avec la distance N-gram entre 'Refuge de montagne' et 'Refuge de montagne des Pyrénées' est de 0,82 ; cette fonction renvoie 1 quand deux chaînes de caractères sont identiques. En revanche, la même comparaison en utilisant la distance de Levenshtein retourne 12 ; cette fonction renvoie 0 quand deux chaînes de caractères sont identiques et 19 quand elles sont complètement différentes.

Chaque type de propriété et type de relation DBpedia trouvé et sélectionné par le contributeur est respectivement initialisé, sous la forme d'un type de propriété et un type de relation dans le vocabulaire construit pour le moment. Le type de propriété est associé au type de *feature* créé pendant la première étape du processus. Le type de relation est également associé à ce type de *feature* de même qu'à un nouveau type de *feature* créé, si nécessaire, par le processus grâce à DBpedia. Autrement dit, un nouveau type de feature peut être créé si nécessaire à partir d'une classe de l'ontologie DBpedia. Chaque type de propriété et type de relation initialisé possède un nom, et contient aussi l'information détaillée sur la source afin de permettre au contributeur de connaître l'origine de cet élément ainsi que comment il a été retrouvé. Cette information d'origine correspond au mot-clé utilisé dans la recherche, au nom du type de propriété dans DBpedia, le domaine et le *range* du type de propriété correspondant. Par exemple, le contributeur à partir du mot-clé `aéroport` trouve et sélectionne le type de propriété `longueur de piste` et le type de relation `plate-forme de correspondance`[56] provenant de l'ontologie DBpedia. En plus, le système propose au contributeur de créer un type

[56] Par exemple, les aéroports de Paris-Charles-de-Gaulle et Paris-Orly sont les plates-formes de correspondance de la compagnie aérienne Air France car elle concentre la plus grande partie de ses activités de gestion et où elle assure la maintenance de ses avions dans ces aéroports.

de *feature* additionnel `Compagnie aérienne` afin de pouvoir instancier le type de relation `plate-forme de correspondance` avec la classe `Aéroport`. La Figure 50 illustre les éléments de vocabulaire choisis pour l'instant par notre contributeur[57].

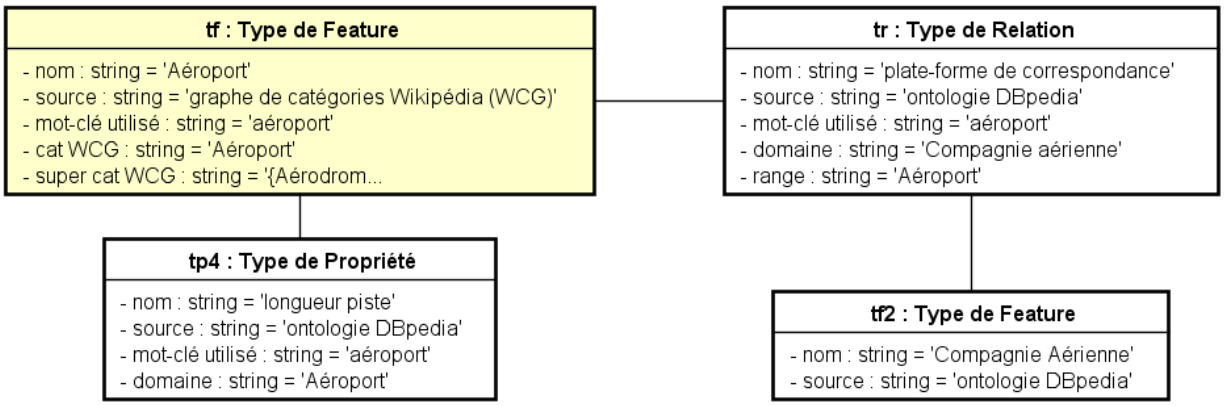



**Figure 50 : un type de propriété et un type de relation du vocabulaire construits à partir du mot aéroport (résultat issu de la mise en œuvre décrite en chapitre 4)**

## 2.3 Extraction d'éléments WordNet : types de relations pour le vocabulaire

Afin d'extraire des types de relations, nous avons d'abord exploré la structure du WCG. Néanmoins, les types de relations lexicales entre les catégories sont en majorité des hyponymes et dans de rares cas de méronymie (Hecht et Raubal 2008). Comme source de type experte, nous utilisons la ressource linguistique WordNet (Miller 1995). Il est possible d'extraire précisément des types de relations lexicales : hyperonymie / hyponymie, méronymie / holonymie. Une relation d'hyperonymie est une relation entre deux termes distinguant le terme le plus général de celui plus spécifique. La relation d'hyponymie est le contraire que celle d'hyperonymie, en distinguant le terme le plus spécifique de celui le plus général. Une relation de méronymie distingue la partie du tout. La relation d' holonymie est le contraire que celle de méronymie, en distinguant le tout de la partie. Les relations hyperonymie / hyponymie et méronymie / holonymie, correspondant souvent à des types de relations de spécialisation et de composition, respectivement. En effet, nous pouvons trouver dans WordNet des types de relations comme « un trottoir est une composante d'une route ». La Figure 51 illustre les *éléments WordNet* extraits et leurs correspondances avec les *éléments de vocabulaire* à construire.

[57] A cause de l'espace, seulement les nouveaux éléments du vocabulaire construit sont affichés.

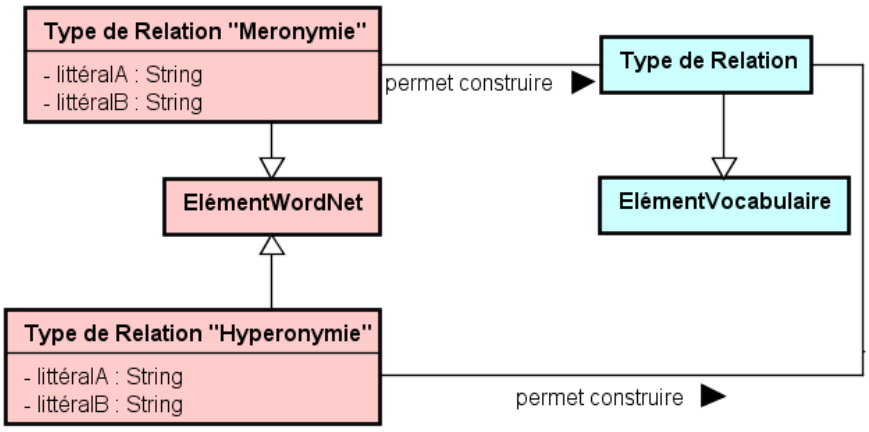



**Figure 51 : les éléments WordNet extraits et leurs correspondances avec les éléments de vocabulaire à construire**

### 2.3.1 Description de WordNet

WordNet est une ressource largement utilisée pour la découverte de relations lexicales entre des termes. Cette ressource est intégrée dans des dictionnaires en ligne comme un support linguistique (ex : The Free Dictionnary). Nous avons utilisé une version européenne de WordNet appelé EuroWordNet6. Cette version contient environs 22.500 littéraux entre des verbes et des noms, et a été construite et vérifiée manuellement par des experts.

### 2.3.2 Extraction des types de relations

A partir du mot-clé et des catégories (et des sous-catégories, si elles sont indiquées) trouvés dans la première étape du processus, il est possible d'extraire des types de relations à partir de WordNet dont un des deux littéraux ressemble syntaxiquement à ce mot-clé ou à un de ces catégories (ou sous-catégories), selon la mesure de similarité N-Gram (N=3) (Euzenat et Shvaiko 2007).

Comme précédemment, chaque type de relation WordNet trouvé et sélectionné par le contributeur est respectivement initialisé, sous la forme d'un type de relation dans le vocabulaire construit pour le moment. Le type de relation est associé au type de *feature* sélectionné dans la première étape du processus, de même qu'à un nouveau type de *feature* créé, si nécessaire, par le processus grâce à WordNet. Le type de relation possède un nom, `est un` pour les relations lexicales hyperonymie/hyponymie ou `est partie de` pour les relations lexicales méronymie/holonymie. Il contient aussi l'information détaillée sur la source afin de permettre au contributeur de connaître l'origine de cet élément et comment il a été retrouvé. Cette information d'origine correspond au mot-clé utilisé dans la recherche, au nom du type de relation lexicale dans WordNet, aux littéraux impliqués dans cette relation lexicale. Par exemple, le contributeur à partir du mot-clé `aéroport` trouve et sélectionne le type de relation `est un` provenant de

WordNet entre le type de *feature* créé auparavant `Aéroport` et un nouveau type de *feature* `Aérodrome` créé par le processus. La Figure 52 illustre les éléments de vocabulaire choisis pour l'instant par notre contributeur

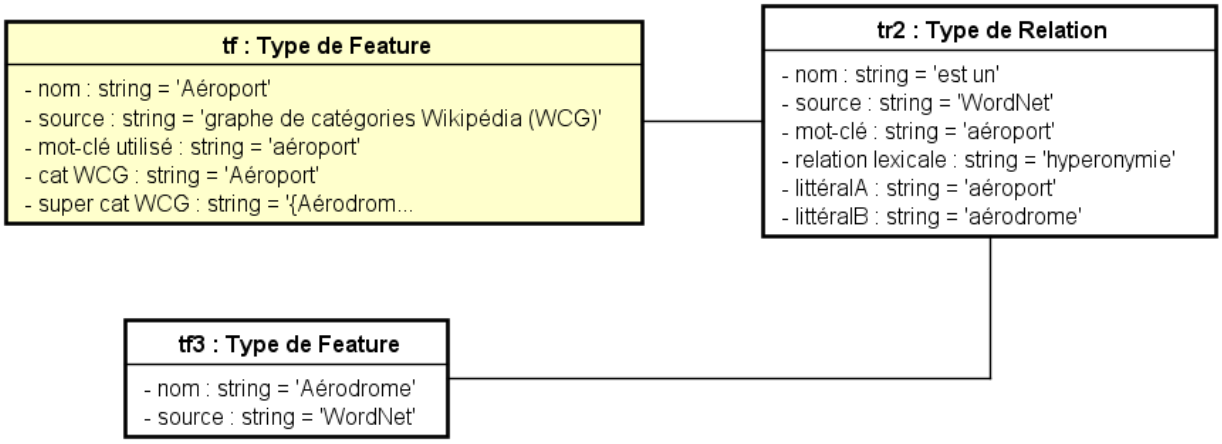



**Figure 52 : type de relation du vocabulaire construit à partir du mot aéroport (résultat issu de la mise en œuvre décrite en chapitre 4)**

## 2.4 Extraction d'éléments IGN : classes pour créer des liens entre le schéma d'un référentiel de données et le vocabulaire

Le but principal de cette étape du processus est d'aider le contributeur à raccrocher ses données au référentiel de données IGN par des types de relations. Plus précisément, le processus propose des types de relations potentiellement pertinents entre des classes du vocabulaire et des classes du référentiel de données. Si aucun type de relation n'est satisfaisant pour le contributeur, il est également possible de réutiliser des types de relations prédéfinis dans le système (ex : `intersecte`), et ainsi de définir une contrainte d'intégrité. Le système peut donc évaluer cette contrainte car les méthodes `évaluer` du type de relation et `corriger` de la contrainte sont implémentées (*cf.* section III.1.3). Pour trouver des classes IGN impliquées dans types de relations, le processus se sert d'une ressource externe de type expert fournie par le Laboratoire COGIT. Une taxonomie de concepts géographiques a été semi-automatiquement dérivée à partir des spécifications en texte libre des bases de données BDTopo® et BDCarto® de l'IGN (Abadie et Mustière 2010). Dans le cadre des mêmes travaux, les auteurs fournissent une version en XML du schéma conceptuel produit de la BDTopo® de l'IGN. La Figure 53 illustre les *éléments IGN* extraits et leurs correspondances avec les *éléments de vocabulaire* à construire.

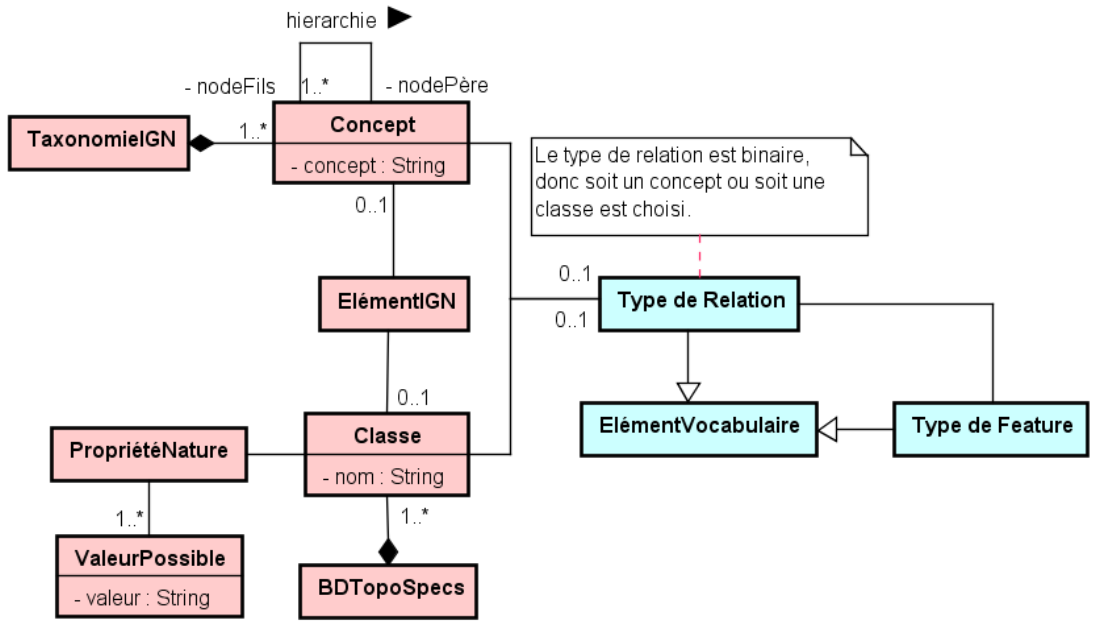



**Figure 53 : les éléments IGN extraits et leurs correspondances avec les éléments de vocabulaire à construire**

### 2.3.1 Description de la taxonomie IGN et de la version formelle des spécifications BDTopo®

La taxonomie de concepts géographiques du COGIT (Abadie et Mustière 2010) est formalisée en OWL et a été construite à partir de l'analyse semi-automatique de documents textuels particuliers : les spécifications des bases de données IGN BDTopo® et BDCarto®. Plus de 700 concepts ont été identifiés et hiérarchisés dans cette taxonomie. La Figure 54 montre un extrait de cette taxonomie géographique à partir du concept `équipement de loisir`.

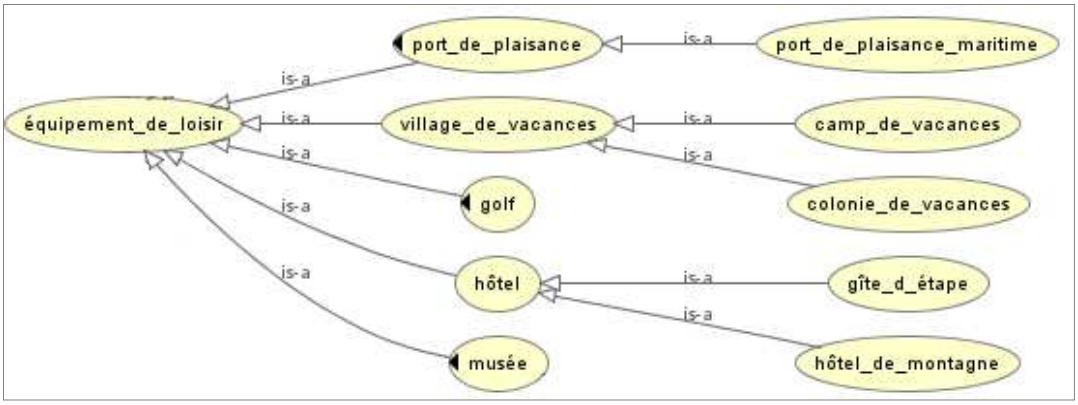



**Figure 54 : un extrait de la taxonomie de concepts géographiques (Abadie et Mustière 2010) générée via l'éditeur d'ontologies OWL Protégé**

De la même manière, un extrait du fichier XML décrivant les spécifications BDTopo® a été produit par les mêmes auteurs et réutilisé par (Kergosien et al. 2010) pour la construction et l'enrichissement automatique d'ontologies à partir de ressources externes. Ces spécifications formelles BDTopo® sont présentées sur la Figure 55.

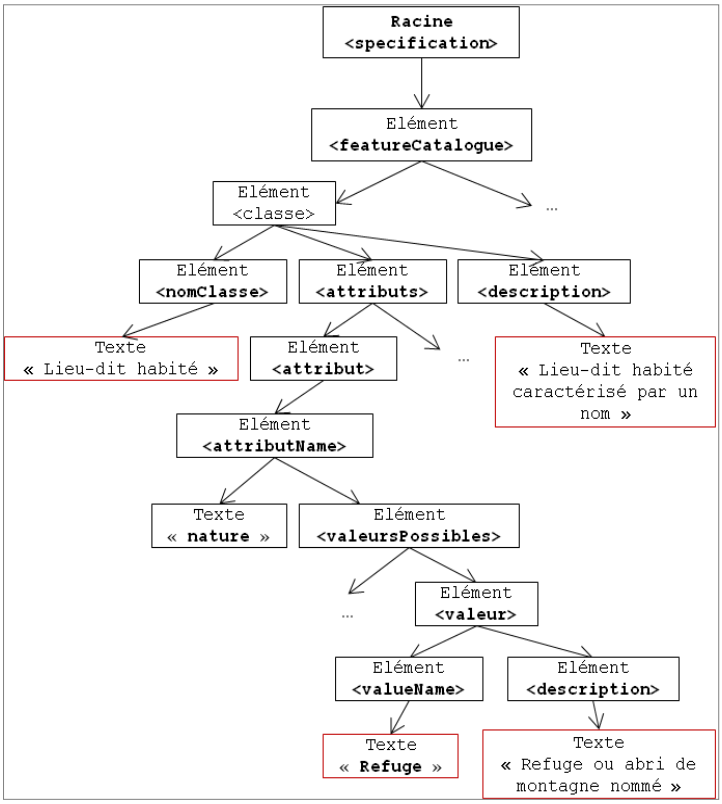



**Figure 55 : extrait de l'arbre XML correspondant à la classe `Lieu-Dit Habité` des spécifications formelles BDTopo®**

### 2.3.2 Extraction des classes IGN afin de créer des types de relations avec les classes du vocabulaire

A partir du mot-clé et des catégories (et des sous-catégories, si elles sont indiquées) trouvées dans la première étape de notre processus, il est possible d'extraire des classes et des concepts IGN. Les concepts IGN sont extraits à partir de l'ontologie IGN en comparant syntaxiquement le mot-clé (aussi les noms des catégories et des sous-catégories correspondantes) et le nom de la classe IGN, selon la mesure de similarité lexicale N-Gram (avec N=3) (Euzenat et Shvaiko 2007). Les classes IGN sont extraites à partir de la version formelle des spécifications de la BDTopo®. Une classe IGN correspondant à un nœud XML `<classe>`, est choisie si ses nœuds XML `Texte` (soulignés en rouge dans la Figure 55) contiennent le mot-clé, les noms des catégories, ou les noms des sous-catégories.

Considérant que le type de relation n'est pas encore connu, le processus utilise ensuite les types de relations trouvés et sélectionnés dans les deux étapes précédentes (extraction d'éléments DBpedia et WordNet). Les noms des classes DBpedia et aussi les littéraux WordNet impliqués dans ces types de relations sont comparés par chaînes de caractère à la liste de classes et des concepts IGN. Le contributeur sélectionne ensuite un type de relation entre une classe ou un concept IGN et une classe du vocabulaire, qui est ajouté à la liste d'éléments du vocabulaire partiel. Le processus peut être capable de trouver un « lien » inconnu entre le type de feature contributeur et la classe IGN. Cependant, il ne peut pas connaitre la nature de cette relation. Le contributeur est donc invité à nommer ce type de relation.

Par exemple, le contributeur à partir du mot-clé `aéroport` trouve grâce au processus qu'un membre du type de relation créé auparavant `est un` correspond à une classe IGN `Aérodrome`. Le processus montre la proposition au contributeur et puis garder ce type de relation ainsi. Il faut noter que le système avait proposé au contributeur de créer un type de *feature* `Aérodrome` dans l'étape précédente, mais le contributeur a explicité qu'il préfère garder le type de relation entre le type de *feature* `Aéroport` et le type de *feature* `Aérodrome` (concept/classe provenant de la source IGN). La Figure 56 illustre ce type de relation choisi par notre contributeur.

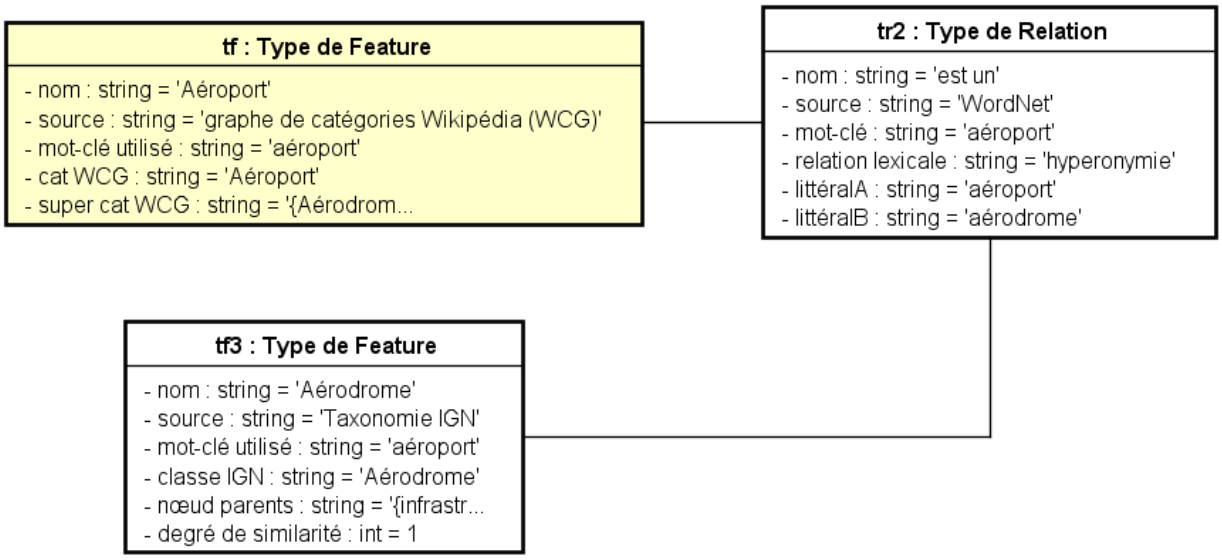



**Figure 56 : type de relation trouvé entre une classe IGN et une classe du vocabulaire à partir du mot aéroport (résultat issu de la mise en œuvre décrite en chapitre 4)**

Pour résumer, le vocabulaire qui peut être construit à partir du mot-clé `aéroport` grâce à notre *processus de construction d'un vocabulaire formel*, est montré sur la Figure 57.

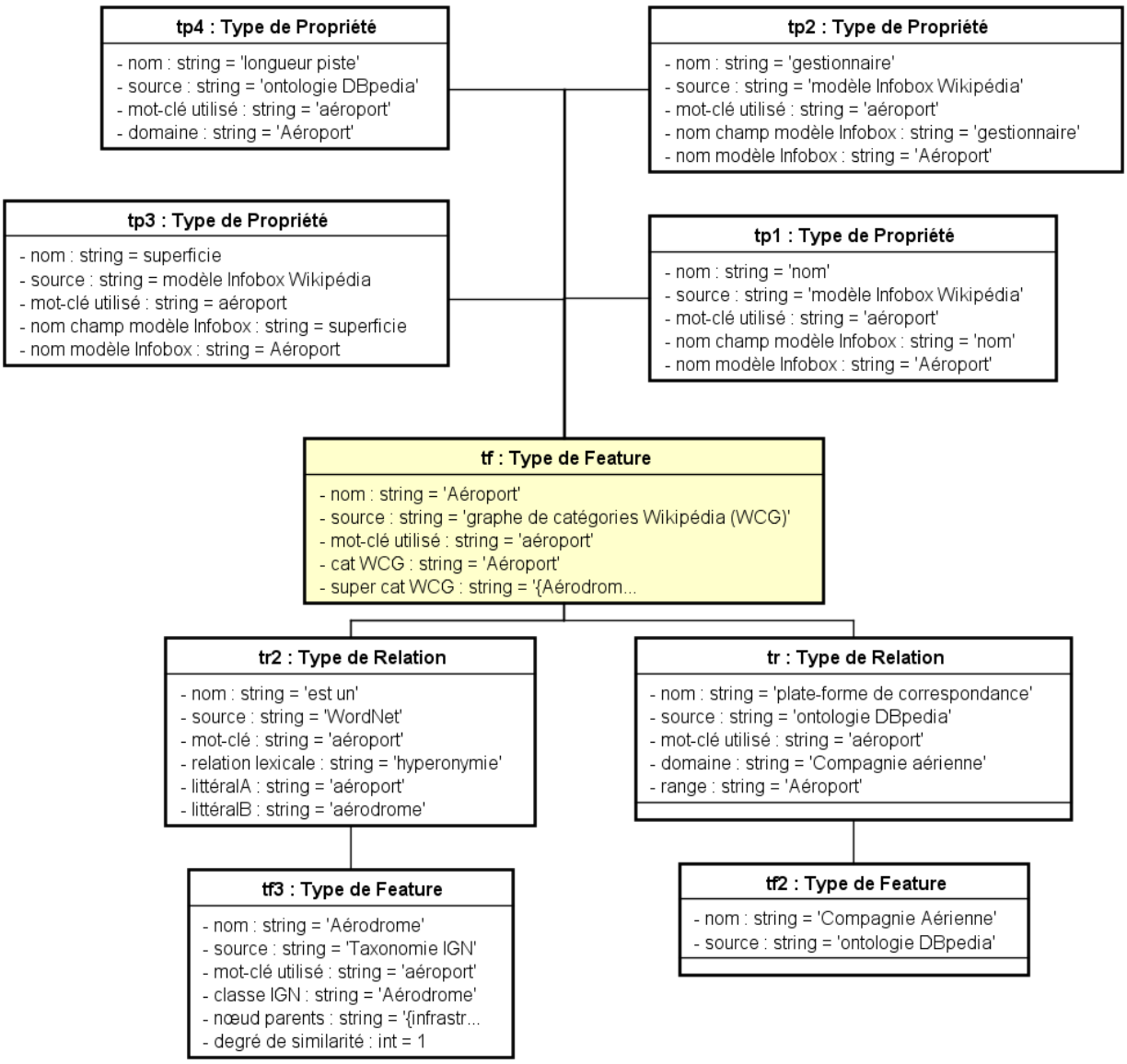



**Figure 57 : vocabulaire construit par notre contributeur grâce au processus proposé à partir du mot-clé aéroport (résultat issu de la mise en œuvre décrite en chapitre 4)**

## 2.5 Extraction de types de relations spatiales à partir des articles Wikipédia

Un élément clé de notre approche est de suggérer au contributeur des instances de types de relations pour l'aider à construire son vocabulaire. Nous avons remarqué que les modèles existants ne possèdent pas certains types de relations topologiques et d'orientation qui peuvent être intéressants pour définir des contraintes d'intégrité pour la gestion de la cohérence. Pour cette raison, nous avons conçu une *méthode d'extraction de types de relations spatiales à partir des articles Wikipédia* qui représentent des entités géographiques (ex : l'article sur La Tour

Eiffel). Le choix de Wikipédia est cohérent avec notre *processus de construction d'un vocabulaire formel* qui utilise l'encyclopédie en ligne pour extraire des catégories et des modèles infobox. Plus précisément, il est nécessaire d'avoir la même orthographe entre le nom de la catégorie d'un article et la catégorie trouvée par notre processus (voir section III.2.1). Cette méthode nous permet également d'explorer la capacité d'une source collaborative comme Wikipédia de fournir ce type d'information et de connaitre ce que les contributeurs de Wikipédia saisissent car les liens entre les projets du libre sont en général forts. Ce travail est inspiré du travail de Massiot et al. (2011) sur la caractérisation d'objets géographiques en fouillant le contenu de Wikipédia.

Des informations précieuses concernant la description d'une entité géographique et ses relations spatiales avec d'autres entités peuvent être fréquemment trouvées dans le premier paragraphe d'un article. En effet, le contenu du premier paragraphe de Wikipédia au contraire du reste du texte de l'article représente une information assez précise (Zesch et Gurevych 2010). Par exemple, le premier paragraphe de l'article `Fontaine du bassin Soufflot` appartenant à la catégorie `Fontaine du 6ème arrondissement de Paris` contient la phrase suivante en wikicode : `La '''fontaine du bassin Soufflot''', appelée aussi le Bassin Pastoral, est située dans le [[6e arrondissement de Paris|6 {{e}} arrondissement]] de [[Paris]], sur la [[place Edmond-Rostand|place Edmond Rostand]] en face de la [[Rue Soufflot]], qui emmène au [[Panthéon]].` Ici, les relations spatiales sont exprimées en texte libre, et les entités impliquées sont indiquées sous la forme des wiki-liens. Il est possible d'exploiter certaines de ces informations en profitant des wiki-liens, comme textes d'ancrages en wikicode, afin de repérer des relations spatiales.

La *méthode d'extraction de types de relations spatiales à partir des articles Wikipédia* repère, au niveau des instances, des relations spatiales dans les textes des articles Wikipédia. Les textes utilisés correspondent à des articles qui décrivent des entités appartenant à un groupe choisi de types de *feature* (ex : fontaines parisiens). La méthode calcule ensuite les fréquences d'apparition de ces relations entre pairs d'entités appartenant aux types de *features* choisis. De cette façon, il est possible de déterminer quels sont les types de relations les plus fréquemment utilisés et sur quels types de features sont-ils définis. En opposition à d'autres étapes de notre processus, l'extraction ne peut pas être réalisée à la volée. Il faut auparavant, appliquer cette méthode pour chaque type de feature initialisé dans le catalogue. Ensuite, il est possible d'intégrer les types de relations extraits dans le catalogue, afin qu'ils puissent être suggérés au contributeur. L'extracteur a été implémenté par Nassima Chenachena dans le cadre de son stage de master 2 recherche en informatique. La Figure 58 illustre la *méthode d'extraction de types de relations spatiales à partir des articles Wikipédia* qui est décrite ensuite en détail.

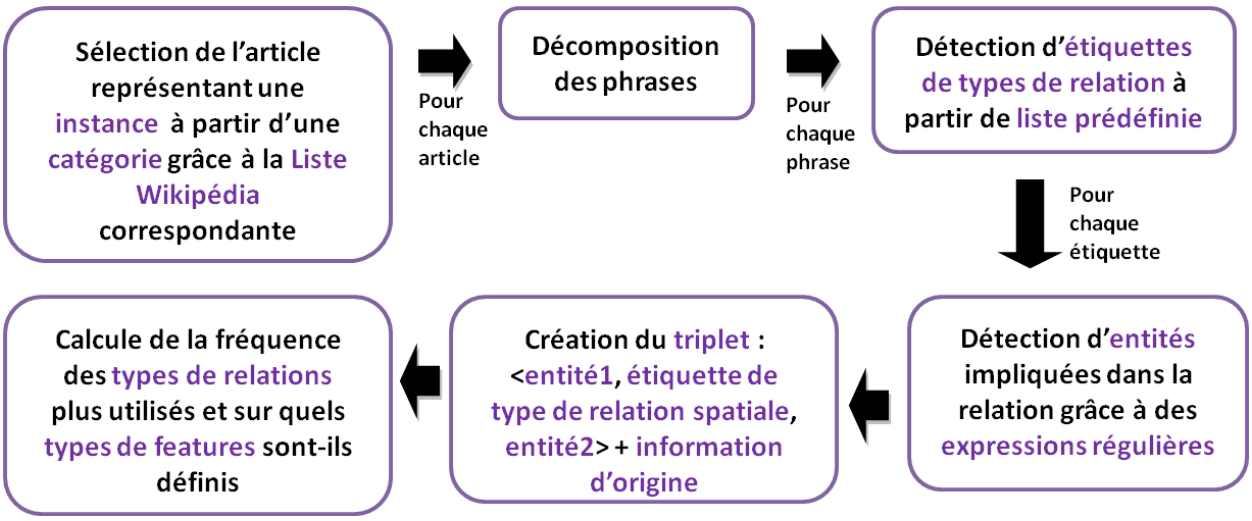



**Figure 58 : méthode d'extraction de types de relations spatiales à partir du texte des articles Wikipédia**

### 2.5.1 Identification des instances géographiques dans Wikipédia

Pour faciliter la tâche d'identifier des entités géographiques, nous nous sommes servis d'un élément d'organisation du contenu de Wikipédia : les pages de type `Liste`. Une page `Liste` concerne une catégorie Wikipédia correspondante à un « concept localisé » et contient les noms des articles correspondant à des entités dans cette catégorie. Par exemple, la page appelée `Liste des fontaines de Paris 4ème arrondissement` liste les noms et les liens Wikipédia vers les articles correspondants à des fontaines du 4ème arrondissement de Paris. La Figure 59 montre un item dans cette page concernant la `fontaine Millénaire`, son nom, sa localisation, ses coordonnées géographiques, ses créateurs, sa date de construction et une image). La Figure 60 montre la version équivalente en wikicode de cet item sur ce même page.

| | | | | | |
|---|---|---|---|---|---|
| Fontaine Millénaire | Parvis Notre-Dame - place Jean-Paul-II | 48° 51' 13" N 2° 20' 57" E | Radi Designers | 2000 |  |

**Figure 59 : description de la `fontaine Millénaire` dans la page `Liste des fontaines de Paris 4ème arrondissement`**

```
|-
| {{sort|millenaire|[[Fontaine Millénaire]]}}
| {{sort|notre-dame parvis|[[Parvis Notre-Dame - place Jean-Paul-II]]}}
| {{coord|48.853619|2.34905|format=dms|region:FR-75|name=Fontaine Millénaire}}
| [[Radi Designers]]
| 2000
| [[Fichier:Fontaine publique paris notre dame 2006.jpg|100px]]
|-
```

**Figure 60 : description en wikicode de la `fontaine Millénaire` dans la page `Liste des fontaines de Paris 4ème arrondissement`**

Afin de sélectionner des instances, l'extracteur sélectionne des articles Wikipédia appartenant à la catégorie trouvée dans la première étape du processus.

### 2.5.2 Extraction des relations spatiales et construction des triplets

Une fois les entités identifiées, la méthode effectue, pour chaque article, une décomposition simple des phrases trouvées dans le premier paragraphe. Il repère des relations spatiales binaires remarquables dans chaque phrase à partir d'une liste prédéfinie d'étiquettes en texte libre de types de relations spatiales (Miron et al. 2007), par exemple, `sur la` et `en face de la`. Nous faisons une hypothèse importante qui nous illustrons avec l'exemple précédent du texte de la `Fontaine du bassin Soufflot` où la méthode peut extraire deux relations. La première entité impliquée dans la relation correspond à l'instance qui est en train d'être analysée, ici `Fontaine du bassin Soufflot`. Ainsi, l'entité `Fontaine du bassin Soufflot` est la première entité des deux relations identifiées. La deuxième entité impliquée dans la relation correspond à un lien interne trouvé dans la première phrase de l'article. Ici, `Place Edmond Rostand` est la deuxième entité de la relation `sur la` et l'entité `Rue Soufflot` est celle de la relation `en face de la`.

Ainsi, l'extracteur construit des triplets sous la forme de `<entité1, étiquette de type de relation spatiale, entité2>`. La méthode produit également un graphe orienté afin d'explorer visuellement les relations trouvées. Un nœud de ce graphe représente une entité d'une des catégories choisies et son arc sortant représente le terme correspondant à la relation spatiale trouvée dans le texte de l'article concernant l'entité. Cet arc est dirigé vers une autre entité appartenant à une des catégories choisie et référencée dans le texte de l'article sous forme de lien interne. D'autre part, le triplet est accompagné d'une information d'origine nous permettant de vérifier comment la relation spatiale a été trouvée. Cette information est composée du nom de l'article et de la phrase dans laquelle le triplet a été trouvé. L'extracteur est aussi capable de traiter d'autres paragraphes des articles Wikipédia. En effet, nous avons remarqué des sections potentiellement pertinentes dans les articles sur une instance géographique précisant des informations sur sa localisation, par exemple, `services et`

`accès` et `emplacement`. Dans ce cas, l'information d'origine contient également le titre de la section dans lequel le triplet a été trouvé.

### 2.5.3 Extraction des types de relations spatiales

Ensuite, la méthode calcule la fréquence d'apparition des relations spatiales trouvées auparavant et sur quels *types de features* à partir des triplets. Elle construit un histogramme nous permettant de détecter les types de relations trouvés et sur quels types de features sont-ils définis. Ensuite, nous stockons dans le *catalogue d'éléments de vocabulaire formel*.

# 3 Une stratégie pour l'intégration de contributions dans un contenu géographique collaboratif

Cette section explique notre stratégie d'intégration des contributions dans un contenu géographique collaboratif. D'abord, nous présentons notre modèle pour l'édition collaborative et la gestion de la cohérence de ce contenu, qui est l'extension du modèle présenté précédemment dans la Figure 34. La deuxième partie de cette section détaille nos stratégies d'*évaluation* et de *réconciliation* de contributions provenant de différents contributeurs afin de les intégrer d'une façon cohérente dans un contenu géographique collaboratif.

## 3.1 Le modèle d'éditions pour l'édition collaborative et la gestion de la cohérence

Cette sous-section détaille notre modèle, illustré sur la Figure 61, pour l'édition collaborative et la gestion de la cohérence.

Une édition est une opération effectuée sur un élément du vocabulaire (ex : créer un type de *feature* du vocabulaire) ou sur les `GéoDonnées` : `Feature`, `Propriété`, `Relation`, (ex : modifier la propriété d'un objet ou d'une relation). Nous nous sommes appuyés du modèle d'édition proposé par (Michaux et al. 2011). Les types d'édition précisés dans notre modèle sont présentés en Figure 62 et détaillés ensuite.

`CréerFeature` (`typeFeature`): crée un nouvel objet dans le contenu, appartenant à la classe du vocabulaire `typeFeature`. Si le paramètre est vide, l'objet appartiendra à la classe générique `Thing`. Un identifiant temporaire est assigné à l'objet par le client (un entier négatif), puis le serveur crée un nouvel identifiant global (un entier non-négatif).

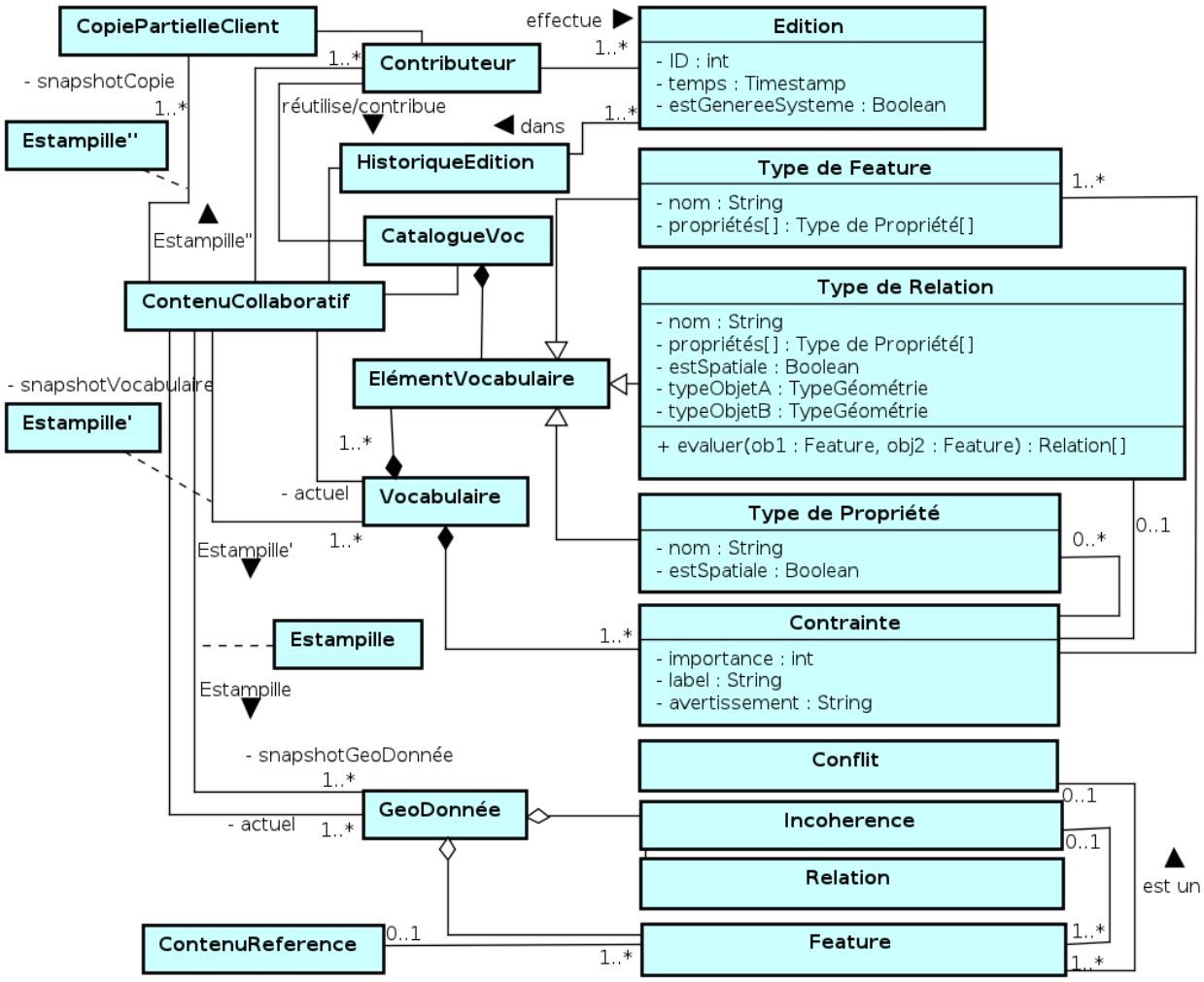



**Figure 61 : modèle général proposé pour l'édition collaborative et la gestion de la cohérence d'un contenu géographique collaboratif (quelques liens sont omis pour la lisibilité)**

`CréerRelation` (`typeRelation`, `feature1`, `feature2`), crée une nouvelle relation binaire explicite dans le contenu, appartenant au type de relation `typeRelation` du vocabulaire entre les objets `feature1` et `feature2`.

`SupprimerFeature` (`feature`), supprime du contenu l'objet `feature`.

`SupprimerRelation` (`relation`), supprime du contenu la relation `relation`.

`ModifierValeurPropriétéFeature` (`feature, propriété, nouvelleValeur`), modifie la valeur de la propriété `propriété` du `feature` en assignant la valeur `nouvelleValeur`.

`ModifierValeurPropriétéRelation (relation, propriété, nouvelleValeur)`, modifie la valeur de la propriété `propriété` de `relation` en assignant la valeur `nouvelleValeur`.

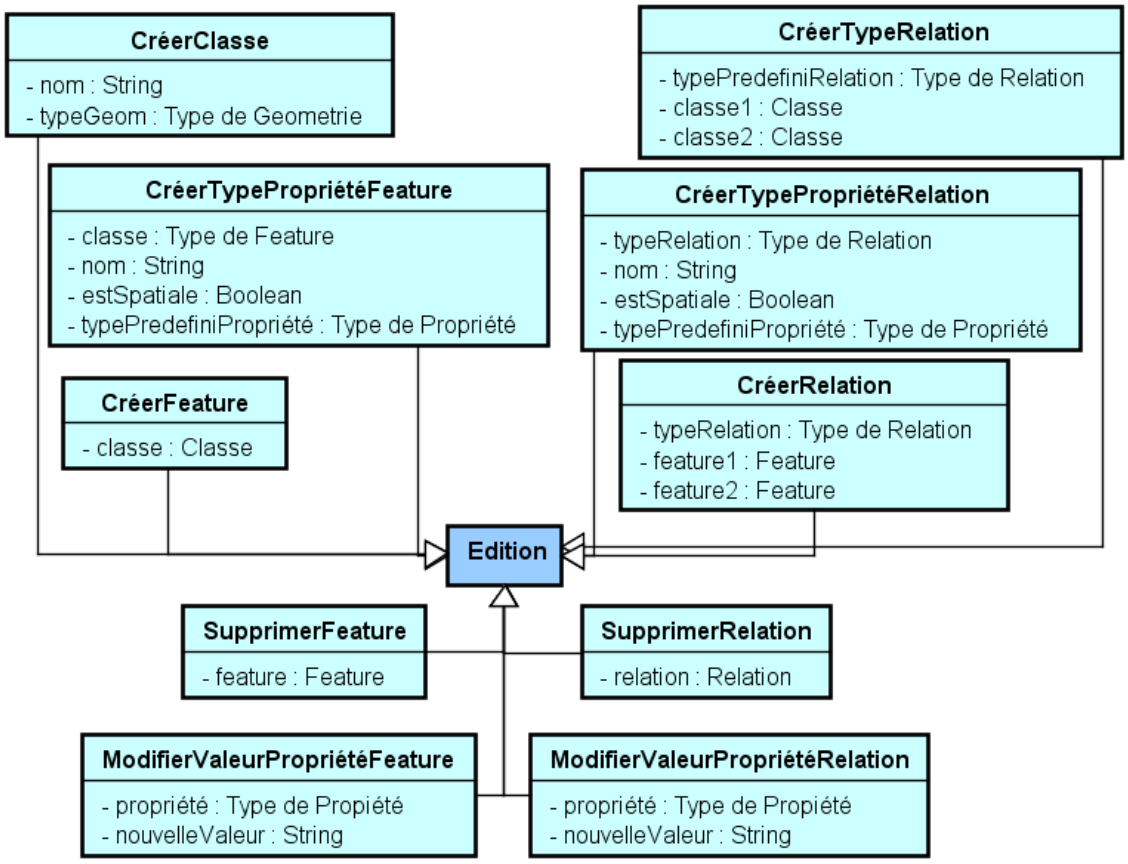



**Figure 62 : types d'éditions possibles**

Les types d'édition visant à changer une version du vocabulaire sont présentés ci-dessous:

`CréerTypeFeature (nom, typeGeometrie)`: crée un nouveau type de *feature*, élément du vocabulaire, avec son `nom` et son type de géométrie `typeGeometrie` dont les valeurs possible sont `polygone`, `polyligne`, ou `ponctuel`.

`CréerTypeRelation (typePredefiniRelation, typeFeature1, typeFeature2)`, instancie un type de relation binaire dans le vocabulaire entre les classes `typeFeature1` et `typeFeature2`, à partir d'un type prédéfinie de relation `typePredefiniRelation` (ex : `franchit` ou `intersect`).

`CréerTypePropriétéFeature (typePredefiniPropriété, typeFeature, nom, estSpatiale)`, instancie un type de propriété `typePredefiniPropriété` dans le

vocabulaire au type de *feature* `typeFeature`. Le type de propriété possède un `nom` et un caractère ou spatiale (ou pas) `estSpatiale`, et peut correspondre à un type prédéfinie de propriété `typePredefiniPropriété` (ex : `forme` ou `nom`) si le premier paramètre est vide.

`CréerTypePropriétéRelation (typePredefiniPropriété, typeRelation, nom, estSpatiale)`, instancie un type de propriété `typePredefiniPropriété` dans le vocabulaire au type de relation `typeRelation`. Le type de propriété possède un `nom` et un caractère ou spatiale (ou pas) `estSpatiale`, et peut correspondre à un type prédéfinie de propriété `typePredefiniPropriété` (ex : `surface d'intersection` pour le type de relation `longer`) si le premier paramètre est vide.

Une `Edition` est appliquée à une version `i` d'un élément ou d'une donnée (sauf pour les créations) et vise à produire une nouvelle version `i+1` de tel élément ou de telle donnée. Une édition est généralement effectuée par un `Contributeur` sur sa `Copie Locale du Contenu`, et elle sera appliquée sur le `Contenu Collaboratif`. Elle représente ce qui a changé entre sa `Copie Locale du Contenu` et le `Contenu Collaboratif`, c'est-à-dire, le différentiel du contenu. Un différentiel permet d'isoler et de décrire les évolutions intervenues entre deux versions d'un contenu (Badard 2000). Une édition peut aussi être créée par le système pour réparer une incohérence (ex : déplacer un bâtiment qui chevauche une route) et proposée au contributeur.

Le versionnement des `GéoDonnées`, des `Vocabulaires`, et des `Copies Partielles du Contenu` des contributeurs, est géré grâce à un `Historique d'Editions`. Il contient les éditions appliquées tout au long du temps (pendant un période limitée). L'historique d'éditions stocke les éditions validées par le système (les éditions correctrices) de même que les éditions refusées par le système (les éditions proposées par le contributeur introduisant une possible incohérence). Grâce à ces `Historiques d'éditions`, il est plus facile de rechercher des évolutions et de répondre à des questions comme : « Quel est l'objet dont la géométrie est souvent modifiée par les contributeurs ? ». L'avantage de stocker les « mauvaises » éditions est de permettre l'analyse *a posteriori* des erreurs fréquentes commises par les contributeurs et de leur offrir une aide ciblée. De plus, nous avons inclus dans le modèle une manière de pouvoir toujours reconstruire une version déterminée de l'objet. Une "photo des données à un instant précis" est stockée régulièrement dans la base de données (classiquement connu comme *snapshot* dans les systèmes classiques de sauvegarde des fichiers). Les copies subséquentes contiendront seulement ce qui a changé par rapport à la dernière copie intégrale, c'est-à-dire un sous-ensemble de l'historique d'éditions. Ce mécanisme permet de construire une version de l'objet à partir du dernier *snapshot* en appliquant les éditions effectuées depuis cette date là.

## 3.2 Les stratégies d'intégration de contributions

Cette section détaille nos stratégies d'évaluation et de réconciliation de contributions provenant de différents contributeurs afin de les intégrer d'une façon cohérente dans un contenu

géographique collaboratif. Pour une séquence d'éditions provenant d'un contributeur, la stratégie d'évaluation vise à détecter des incohérences en évaluant les contraintes d'intégrité pour la préservation de relations spatiales et à détecter des conflits d'éditions concurrentes en vérifiant des dépendances entre les propriétés. Elle propose également des éditions correctrices afin de résoudre les incohérences. Pour deux séquences d'éditions provenant de différents contributeurs, la stratégie de réconciliation vise à fusionner ces séquences.

### 3.2.1 Evaluation d'une séquence de contributions d'un utilisateur dans le contenu

Cette sous-section présente la stratégie d'intégration d'une séquence d'éditions proposée par un contributeur qui pourrait introduire des incohérences dans le contenu central. Pour détecter les incohérences, cette stratégie évalue dans un premier temps les contraintes potentiellement violées, pour ensuite tenter de résoudre ces incohérences en effectuant des transformations géométriques sur la donnée concernée par la séquence d'éditions. Ces transformations sont proposées sous la forme d'éditions correctrices au contributeur qui doit valider la proposition du système. La détection d'incohérences et la suggestion de corrections est similaire à la proposition de Servigne et al. (2000).

La stratégie est composée des étapes suivantes :

- récupération des contraintes potentiellement violées,
- extraction d'un jeu de données d'évaluation,
- création d'un jeu de données temporaire,
- détection d'incohérences liées à la non-préservation de relations spatiales importantes,
- calcul des éditions correctrices et,
- proposition des corrections afin d'être validées par le contributeur.

Ces étapes sont exprimées en pseudo-code ci-dessous et détaillées ensuite.

```
1: potentViolConstraints <- obtainPotentViolConstraints(editions)
2: geoDataSets <- obtainGeoDataSets(potentViolConstraints)
3: tempGeoDataSets <- createTempGeoDataSets(editions)
4: for all constraint in potentViolConstraints do
5:    conflict <- evaluate(constraint, geoDataSets, tempGeoDataSets)
6:    conflicts.add(conflict)
7: end for
8: for all conflict in conflicts do
9:    editionsSystem<-correct(conflict, tempGeoDataSets, geoDataSets)
10:   editions.addAllItems(editionsSystem)
11: end for
12: return editions; conflicts(descriptions); tempGeoDataSets
```

**Récupération des contraintes potentiellement violées (ligne 1)**

Pour chaque édition qui peut porter sur un ou plusieurs objets parmi la séquence envoyée par le contributeur, le système recherche dans un catalogue prédéfini des `contraintes sur des types de relations` (voir la Figure 38) et des `contraintes de dépendance entres des types de propriétés` (voir la Figure 39), qui peuvent être potentiellement violées par le résultat de cette édition. Il existe deux critères pour considérer qu'une contrainte est potentiellement violée. Le premier critère est de vérifier si la définition de la contrainte concerne une classe (ou un type de relation) de l'objet (ou relation) touché par l'édition courante. Les `contraintes sur des types de relations` sont uniquement concernées par ce critère. Pour cela, la valeur `estSpatiale` des types de relations impliqués dans la contrainte joue un rôle très important. Pour illustrer cela, considérons une édition qui modifie la géométrie d'un objet de type `Abribus` et considérons aussi une contrainte de relation spatiale indiquant `les abribus ne doivent pas intersecter les tronçons de routes`. Cette contrainte concerne un type de relation `intersecte` et deux classes `Abribus` et `Tronçon de route` (voir la représentation d'un type de relation dans la Figure 36). Elle est donc présélectionnée comme une contrainte pouvant être violée par une édition concernant un objet de type `Abribus`. Le deuxième critère consiste à regarder plus précisément le type d'édition (ex : création d'objet, modification de propriété, etc.) et l'attribut duquel dépend la contrainte (sauf pour les créations). Les `contraintes de dépendance entres des types de propriétés` et les `contraintes de dépendance entres des types de propriétés` sont concernées par ce critère. En considérant l'exemple cité précédemment, le système vérifie que l'édition est une modification de la géométrie de l'objet. Cela veut dire que cette édition peut affecter les relations spatiales entre l'objet concerné par l'édition (l'abribus) et d'autres autour lui (certains tronçons de route). C'est pour cela que la contrainte est sélectionnée dans la liste de contraintes qui devront être évaluées. Comme par défaut toutes les propriétés sont indépendantes, le système vérifie également si les propriétés impliquées dans la liste d'éditions d'entrée sont dépendantes grâce à des `contraintes de dépendance entres des types de propriétés`.

**Extraction d'un jeu de données d'évaluation (ligne 2)**

Les objets nécessaires pour l'évaluation des contraintes pouvant appartenir à un contenu de référence ou à la dernière version du contenu central sont extraits du serveur. En considérant que la définition des contraintes contient les classes impliquées et les types impliqués de relations, l'extraction filtre les instances des classes et les types de relations utilisés dans la définition de la contrainte. Des données routières sont extraites du contenu de référence et des abribus sont extraits du contenu central. L'extraction filtre les objets avec une zone englobant la zone de travail du contributeur (en appliquant une opération `buffer`).

**Création d'un jeu de données temporaire (ligne 3)**

Chaque édition de la séquence soumise par le contributeur est appliquée sur le jeu de données d'évaluation pour former un jeu de données temporaire. De cette manière, il est possible de détecter plusieurs incohérences du même type en évaluant la contrainte une seule fois. Par exemple, deux incohérences correspondant à deux abribus chevauchant un même tronçon de route seront détectés par l'évaluation de la contrainte `les abribus ne doivent pas intersecter les tronçons de routes`. Il est important de respecter la séquence initiale d'éditions afin de ne pas introduire de nouvelles incohérences sur d'autres objets touchés par d'autres éditions.

**Détection d'incohérences (lignes 4-7)**

Les incohérences sont détectées grâce à l'évaluation des contraintes potentiellement violées. Plus précisément, une incohérence est une contrainte violée. La stratégie d'évaluation des contraintes utilise la définition de la contrainte (avec les éditions correspondantes qui font de cette contrainte une candidate à être potentiellement violée), les jeux de données d'évaluation et des jeux de données temporaires. Chaque contrainte potentiellement violée est évaluée par des opérations spatiales sur les objets et les relations affectés.

**Calcul des éditions correctrices (lignes 8-11)**

Les incohérences trouvées dans l'étape précédente sont traités l'un après l'autre. Pour chaque incohérence, le mécanisme de résolution d'incohérences est déclenché. Ce mécanisme considère l'incohérence (avec les objets ou les relations affectés), la contrainte, les jeux de données d'évaluation et temporaire. Une contrainte pointe vers la méthode de résolution `corriger` qui devra être invoquée afin de corriger l'incohérence associé à cette contrainte. Par exemple, la contrainte `les abribus ne doivent pas intersecter les tronçons de routes` définit une méthode `corriger` qui utilise la méthode d'analyse spatiale `déplacer`. Plus précisément, cette méthode considère le type de géométrie définie pour la classe `Abribus` (polygonale) afin de calculer l'orientation de l'abribus par rapport au tronçon de route (une polyligne) et l'aire d'intersection entre les deux objets (intersection entre un polygone et une polyligne). La méthode correctrice produit une édition supplémentaire qui modifie la géométrie de l'objet (un abribus) concernée par l'incohérence. Une fois les éditions correctrices calculées, le jeu de données temporaire est mis à jour en les appliquant. Nous supposons que la stratégie collecte assez d'informations pour les méthodes de transformation afin de ne pas créer des nouvelles incohérences. Nous sommes conscients que c'est une stratégie très optimiste, mais nous pensons que dans le contexte de l'édition collaborative, cette stratégie pourrait aider dans de nombreux cas moins complexes que ceux traités en généralisation cartographique. Une librairie des méthodes d'analyse spatiale résolvant ce type d'incohérences (ex : le déplacement d'un bloc de bâtiments en préservant leur alignement ou encore le calcul de la meilleure direction d'échappement d'un groupe de bâtiments) sont disponibles dans la plate-forme de généralisation cartographique CartAGen (Renard et al. 2011).

**Proposition des corrections à valider par le contributeur (ligne 12)**

A la suite de l'étape précédente, le système envoie au contributeur une proposition de correction des incohérences détectées comprenant :

- les éditions correctrices qui sont ajoutées à la fin de la séquence initiale d'éditions,
- la description en langage naturel des incohérences trouvées,
- et le jeu de données temporaire (où les éditions correctrices ont été appliquées) suggéré comme une nouvelle version du contenu.

Si le contributeur accepte le nouveau contenu suggéré par le serveur, les actions suivantes sont effectuées :

1. Des identifiants globaux sont assignés par le serveur aux objets créés,
2. La copie locale du contenu du client et le contenu central sont mis à jour avec le jeu de données temporaires (où les éditions correctrices ont été appliquées),
3. La séquence initiale d'éditions suivie des éditions correctrices sont stockées dans l'historique d'éditions.

Suite à l'incohérence de l'abribus qui se superpose à la route, le serveur propose au client une correction concernant la description de l'incohérence détectée en langage naturel et l'édition correctrice proposée par le serveur.

### 3.2.2 Réconciliation de deux séquences de contributions d'utilisateurs différents dans le contenu

La stratégie de réconciliation vise à fusionner deux séquences indépendantes d'éditions, proposées par deux contributeurs différents. Cette stratégie est composée des étapes suivantes :

- regroupement des éditions des deux listes en des groupes d'éditions dépendantes,
- fusion des créations d'objets ou de relations et,
- fusion des modifications de propriétés. Ces étapes sont exprimées en pseudo-code ci-dessous et détaillées ensuite.

```
1: edClusters <- clusterEditions(edsA, edsB)
2: for all cluster in edClusters do
3:    edsA' <- cluster ∩ edsA
4:    edsB' <- cluster ∩ edsB
5:    createFeatEds <- mergeNewFeatIds(edsA',edsB',linksFeatures)
6:    fusedEds.addAllItems(createFeatEds)
7:    modifyFeatEds <- mergeFeatModifs(createFeatEds,edsA',edsB')
8:    fusedEds.addAllItems(modifyFeatEds)
9:    finalEds <- reconcileEditionsOneContributor(fusedEds)
10: end for
11: return finalEds
```

**Regroupement des éditions en groupes d'éditions dépendantes (ligne 1-2)**

Comme précédemment, deux critères sont utilisés pour déterminer que deux éditions sont dépendantes ou non : les objets concernés sont proches dans l'espace (distance entre leurs géométries inférieure à un paramètre prédéfini : nous avons testé la valeur 200m) et il existe une ou plusieurs contraintes relatives à leurs classes. Autrement dit, un groupe d'éditions est dépendante si les *features* concernés par ces éditions sont proches dans l'espace et s'il existe au moins une relation qui doit être préservée entre les *features* concernés. Selon ces deux critères, un algorithme de classification ascendante hiérarchique est utilisé sur l'ensemble des éditions (indépendamment des listes auxquelles elles appartiennent) pour construire les groupes. La Figure 63 illustre un exemple de deux groupes d'éditions dépendantes calculés selon les deux critères. Dans cet exemple, un contributeur B saisit trois bâtiments et deux abribus et un contributeur A saisit un abribus sur la même version du contenu. Les éditions sont classées selon le premier critère : leur proximité, et puis reclassées selon le deuxième critère : les relations à préserver. Ici, les relations à préserver sont la non-superposition entre les abribus et entre les bâtiments, et aussi que l'abribus devrait normalement être placé entre la route et le bâtiment. De cette manière, nous trouvons les deux groupes #1 et #2 sur cette figure.

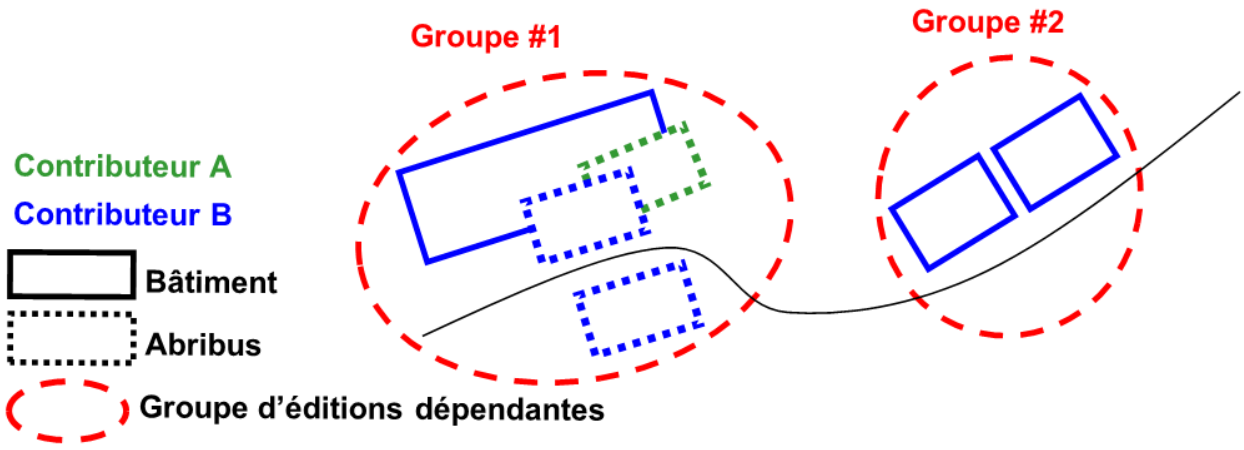



**Figure 63 : groupes d'éditions dépendantes en rouge**

**Fusion des créations d'objets ou des créations de relations (ligne 5-6)**

Le but de cette phase est de résoudre les conflits de duplication liés à la création par chaque contributeur d'un même objet ou d'une même relation. Dans un premier temps, deux sous-listes sont créées : les créations d'un des contributeurs et les créations de l'autre contributeur. A partir de chacune, une autre liste d'objets sont est construite contenant les objets tels qu'ils ont été créés puis modifiés par chaque contributeur. Cela permet de récréer les objets tels qu'ils sont été saisis par le contributeur. La même action est effectuée pour les relations. Ces listes d'objets sont alors appariées. En l'absence d'un algorithme générique d'appariement efficace,

un algorithme dédié aux objets de la classe `Abribus` a été implémenté dans le prototype actuel. Cet algorithme tient compte de leur position, de leurs relations spatiales avec des routes et de leurs informations attributaires comme `lignes de bus` et les `directions des lignes`. Lorsque des objets sont appariés, les créations sont fusionnées par l'attribution d'un même identifiant global à l'objet créé dans les deux listes. Les créations sont mises dans la liste finale d'éditions fusionnées et enlevées des listes initiales.

**Fusion des modifications de propriétés (ligne 7-8)**

Les identifiants des objets sont utilisés pour identifier dans chaque liste les modifications apportées aux mêmes propriétés des mêmes objets. Ensuite, les conflits d'édition de propriétés dépendantes sont détectés pour chaque paire d'éditions ainsi trouvée grâce à l'évaluation des contraintes explicitant la dépendance des types de propriétés. Le système décide que l'édition faite par le contributeur « le plus productif », c'est-à-dire celui qui a fait le plus d'éditions sur l'objet ou sur la relation concernée. Cette édition est ensuite conservée dans la liste d'éditions fusionnées et les deux éditions sont enlevées des listes initiales. Enfin, les éditions restantes dans les deux listes initiales sont mises dans la liste des éditions fusionnées.

**Evaluation d'une séquence d'éditions (ligne 9)**

Les éditions restantes dans les deux listes initiales sont mises dans la liste des éditions fusionnées. Ensuite, la stratégie d'*évaluation d'une séquence de contributions d'un utilisateur* est invoquée sur la liste des éditions fusionnées afin de détecter des incohérences et de proposer des éditions correctrices.

# 4 Analyse comparatif du modèle BDUni de l'IGN et de notre modèle pour l'édition collaborative des données géographiques

Cette section présente une analyse comparative du modèle BDUni utilisé en production à l'IGN et notre modèle, concernant l'aspect des évolutions du contenu. Nous avons réalisé auparavant plusieurs entretiens avec les experts du Service du Développement à l'IGN. Ces discussions nous ont permit d'améliorer et étudier les apports de notre modèle dans le contexte de l'IGN. Ensuite, Nous présentons la manière dont chaque modèle gère les évolutions du contenu. Pour illustrer la façon dont les évolutions sont stockées dans chaque modèle, nous utilisons un exemple réel sur le thème adresse.

## 4.1 Le modèle BDUni

En section I.2.3, nous avons présenté la structure de la base de données relationnelle BDUni de l'IGN. En particulier, la Figure 64 illustre le schéma logique de la BDUni avec le thème des adresses.

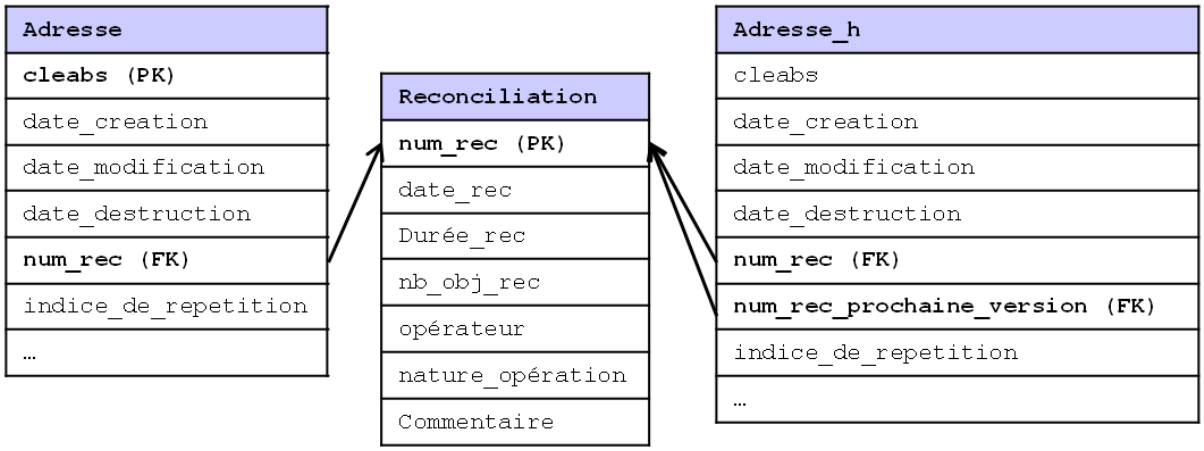



**Figure 64 : extrait du schéma logique de la BDUni concernant les adresses**

Pour rappel, la réconciliation à l'IGN est un processus visant à garder l'intégrité de la BDUni et déclenché par l'opérateur via son client qui comprend deux phases. D'abord, le serveur transfère vers le client les objets étant dans la zone de réconciliation (et autour) qui ont été modifiés auparavant dans la BDUni par d'autres opérateurs. Ensuite, le client transfère vers le serveur, les objets modifiés par lui-même se trouvant dans la zone de réconciliation. L'information sur la réconciliation se trouve dans la table `Reconciliation`. Elle contient le champ `nature_opération` qui décrit le type de changement que l'opérateur souhaite appliquer à ce groupe d'objets. Les valeurs possibles de ce champ sont : `adressage`, `correction`, `documentaire`, `enrichissement`, `MAJ_GE` (mise à jour à grande échelle), `MAJ_ME` (mise à jour à moyenne échelle). L'opérateur utilise aussi le champ `commentaire` pour expliquer en texte libre ses changements, par exemple, `'Ecriture en toutes lettres de l'indice de repetition'`.

L'historique des objets est organisé de façon que chaque nouvelle version `N` d'un objet adresse soit ajoutée dans la table `Objet` et la version précédente `N-1` transférée vers la table `Objet_h`. Cette structure est conforme à celle dans d'autres outils comme les moteurs de Wiki. Par exemple, MediaWiki définit des tables `Page` et `Revision`[58]. De la même manière, la BD Open Street Map contient des tables `current_nodes`, `history_nodes`, `current_nodes _tags`, `history_nodes_tags`[59]. Chaque version d'un objet garde le lien vers le `numéro de`

[58] http://www.mediawiki.org/wiki/Manual:Database_layout

[59] http://wiki.openstreetmap.org/wiki/Database

réconciliation du processus qui l'a générée. Une version ancienne d'un objet garde également le lien vers le numéro de réconciliation correspondant à une révision plus récente de l'objet (num_rec_ prochaine_version).

Les Tableau 3, Tableau 4, et Tableau 5 montrent le stockage des informations concernant les évolutions dans la BDUni et les réconciliations IGN. Ils contiennent, respectivement, la dernière version de l'objet adresse avec l'identifiant 'ADRNIVX_0000000 287165726' ('ADR_2' pour simplifier), ses trois versions précédentes, et les quatre processus de réconciliation donnant lieu à toutes ces versions.

| cleabs | num rec | num | indice de repetition | nom voie | geometrie | date modif | méthode |
|---|---|---|---|---|---|---|---|
| ADR_2 | 7795038 | 9 | BIS | AV DU BOIS | POINT (638037.6 6859607.9) | 2012-06-28 16:28:37 | terrain |

**Tableau 3 : extrait de la table Adresse de la BDUni contenant la dernière version de l'objet 'ADR_2' avec certaines de ses attributs**

| cleabs | num rec | num rec prochaine version | num voie | indice de rep | nom voie | geometrie | date modif | méthode |
|---|---|---|---|---|---|---|---|---|
| ADR_2 | 6936925 | 7795038 | 9 | BIS | AV DU BOIS | POINT (638037.6 6859607.9) | 2011-12-29 14:54:34 | Prélocalisé |
| ADR_2 | 6496888 | 6936925 | 9 | B | AV DU BOIS | POINT (638037.6 6859607.9) | 2011-09-15 14:32:11 | Prélocalisé |
| ADR_2 | 6333982 | 6496888 | 9 | B | | POINT (638037.6 6859607.9) | 2011-08-10 18:47:49 | Prélocalisé |

**Tableau 4 : extrait de la table Adresse_h contenant les trois versions précédentes de l'objet 'ADR_2' avec des anciennes valeurs de certaines de ses propriétés ordonnées par ordre descendant par la propriété date modif**

| num rec | date rec | duree rec | nb obj rec | opérateur | nature opération | comment |
|---|---|---|---|---|---|---|
| 7795038 | 2012-06-28 16:28:37 | 11 | 191 | TGuillon | Adressage | |
| 6936925 | 2011-12-29 14:54:34 | 21 | 124 | EDorange-Pattoret | Adressage | Ecriture en toutes lettres indice de rep. |
| 6496888 | 2011-09-15 14:32:11 | 5 | 4 | TGuillon | Adressage | |
| 6333982 | 2011-08-10 18:47:49 | 8 | 200 | TGuillon | Correction | |

**Tableau 5 : extrait de la table Réconciliation contenant les quatre processus de réconciliation donnant lieu à toutes les versions de l'objet 'ADR_2', de la version la plus récente à la version la plus ancienne**

Concernant la gestion de l'historique des données, nous observons qu'il existe des informations redondantes d'une version à une autre sur des propriétés qui n'ont pas évoluées. Nous observons également que le concept de différentiel est implicite. La représentation des objets et de leurs évolutions dans la BDUni ne décrit pas directement ce qui a changé dans le contenu entre deux versions différentes. Plus précisément, nous voudrons savoir "le quoi" (ex : une géométrie ou une propriété thématique) et quel est le type de changement, c'est-à-dire "le comment" (ex : une modification, une suppression). Les valeurs des champs `commentaire` et `nature opération` remplis par l'opérateur ne suffisent pas pour connaître la nature de l'évolution de chaque objet. A moins de recourir systématiquement à des opérations de comparaison des valeurs entre deux versions de l'objet dans la table `Objet_H`, nous ne pouvons pas savoir simplement quelle est la propriété d'un objet et comment celle-ci a évoluée tout au long de son existence dans la base de données.

## 4.2 Le modèle *Coalla*

Notre modèle *Coalla*, présenté en section III.3, est basé sur un modèle d'éditions, c'est-à-dire que les évolutions du contenu sont stockées sous forme d'éditions. La Figure 65 montre un extrait de notre schéma conceptuel illustrant la façon dont les évolutions sont représentées. Il faut principalement retenir les classes `Adresse`, `Edition` et `Historique Edition`. Nous gardons dans un historique d'éditions le différentiel entre `N` et `N-1`. Plus précisément, nous stockons le type d'édition effectué sur un objet (création ou suppression) et la modification au niveau de ses propriétés thématiques ou géométriques. Nous avons inclus dans le modèle une manière de pouvoir toujours reconstruire une version déterminée de l'objet. Une "photo des données à un moment donné" est stockée régulièrement dans la base de données (classiquement connu comme *snapshot* dans les systèmes classiques de sauvegarde des fichiers). Ce mécanisme permet de construire une version de l'objet à partir du dernier *snapshot* en appliquant les éditions effectuées depuis cette date là.

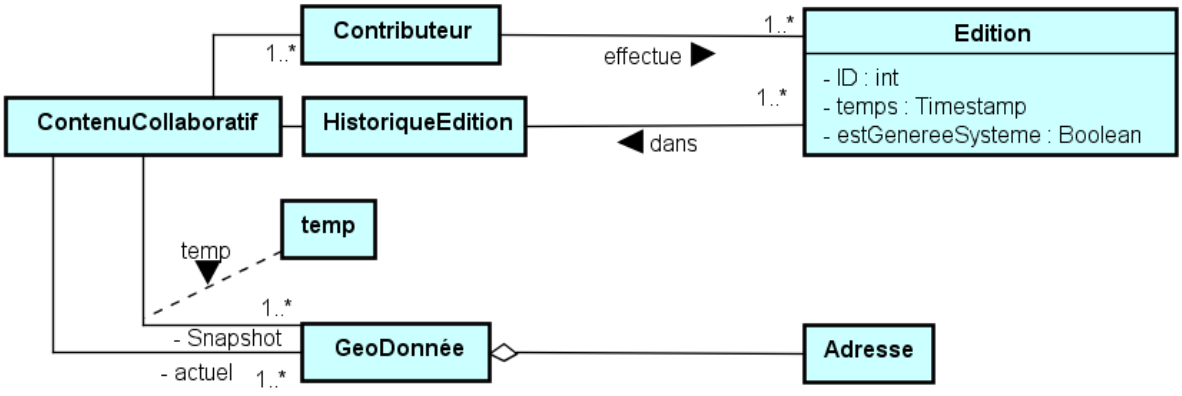



**Figure 65: extrait de notre modèle d'édition collaborative (exemple avec la classe Adresse)**

Les Tableau 6, Tableau 7, et Tableau 8 montrent comment l'objet adresse identifié avec 'Addr_2' de l'exemple précédent, et ses évolutions serait représentés dans notre modèle.

| id objet | géometrie | num | ind de rep | nom voie | cote | méthode |
|---|---|---|---|---|---|---|
| ADR_2 | POINT (638037.6 6859607.9) | 9 | BIS | AV DU BOIS | GAUCHE | terrain |

**Tableau 6 : exemple sur un extrait de la table Adresse montrant la dernière version de l'objet (les propriétés et leurs valeurs)**

| id hist édition | id édition | id objet | type d'édition | id contributeur | timestamp |
|---|---|---|---|---|---|
| H_1 | E_1 | ADR_2 | création | TGuillon | 2011-08-10 18:47:49 |

**Tableau 7 : exemple sur un extrait de la table historique création**

| id hist édition | id édition | id opérateur | timestamp | type d'édition | nom propriété | valeur propriété |
|---|---|---|---|---|---|---|
| H_2 | E_2 | TGuillon | 2011-08-10 18:47:49 | modification | geometrie | POINT (638037.6 6859607.9) |
| H_3 | E_3 | TGuillon | 2011-08-10 18:47:49 | modification | methode | prélocalisé |
| H_4 | E_4 | TGuillon | 2011-08-10 18:47:49 | modification | indice de répetition | B |
| H_5 | E_5 | TGuillon | 2011-08-10 18:47:49 | modification | num | 9 |
| H_6 | E_6 | TGuillon | 2011-09-15 14:32:11 | modification | nom voie | AV DU BOIS |
| H_7 | E_7 | EDorange-Pattoret | 2011-12-29 14:54:34 | modification | indice de répetition | BIS |
| H_8 | E_8 | TGuillon | 2012-06-28 16:28:37 | modification | méthode | terrain |

**Tableau 8 : exemple sur un extrait de la table historique modifications**

Le premier avantage d'utiliser un modèle d'édition est de pouvoir garder un historique des éditions effectuées au contenu. De cette manière, il est plus facile de rechercher des évolutions et de répondre à des questions comme : « Quel est l'objet dont sa géométrie est souvent modifiée et par quels opérateurs? ». Il est également possible de détecter des *patterns* dans l'historique d'éditions montrant des comportements "non-idéaux" à des opérateurs afin de les aider pendant l'édition ou d'éviter l'introduction d'incohérences dans la base de données. Par exemple, en voulant changer la géométrie d'un objet, il arrive souvent qu'un opérateur détruise l'objet et dessine la nouvelle géométrie, il doit alors recopier manuellement tous les attributs de l'ancien objet sur le nouvel objet. Des erreurs de nature humaine de l'opérateur peuvent apparaître, par exemple, mettre des valeurs à vide à un gros lot de propriétés ou effacer beaucoup d'objets.

Le second avantage du modèle d'édition est de permettre l'édition concurrente de zones indépendantes du contenu (une approche optimiste). À l'inverse l'approche pessimiste des systèmes de gestion de bases de données bloque l'accès concurrent au contenu en écriture

lors de l'édition par un contributeur. En partant d'une approche optimiste, il est possible d'accepter des éditions des contributeurs sans restrictions. Ensuite, il est possible de fusionner deux séquences d'éditions, chacune effectuée séparément sur une copie différente du même contenu. De cette manière, il est possible d'identifier les parties indépendantes et de réduire le nombre de faux conflits. Un conflit n'est pas intrinsèquement lié à l'édition d'un même objet par deux opérateurs différents. Par exemple, si des propriétés différentes d'un même objet sont éditées par deux opérateurs différents, il ne devrait exister aucune alerte à l'opérateur et aucune incohérence ne devrait s'introduire dans la base de données.

De plus, les discussions avec les experts nous a permis d'apporter une amélioration à notre modèle, comme montré sur la Figure 66. En particulier, certaines éditions devraient toujours être considérées ensemble, c'est-à-dire, elles sont « inséparables ». Par exemple, la suppression d'un feature F suivie d'une création d'un nouveau feature NF dans le même endroit approximativement. Ainsi, il est possible d'induire que l'intention de l'utilisateur était vraiment de modifier la géométrie du feature F. Ce critère est particulièrement important pour notre stratégie de pendant la création de groupes d'éditions dépendantes.

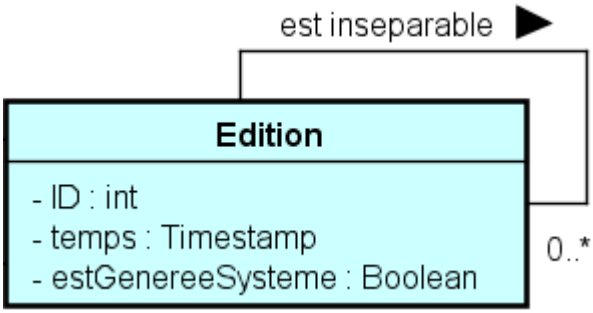



**Figure 66 : modification de la classe Edition de notre modèle après suggestions des experts de l'IGN**

# Chapitre IV
# Mise en œuvre

Ce chapitre présente le prototype qui a été mis en œuvre à partir de notre proposition décrite en chapitre III. Il est relativement difficile de trouver des cas d'applications réels pour valider l'ensemble des composantes de notre proposition, c'est pour cela que nous avons décidé de tester indépendamment certaines composantes clés.

La section 1 présente les spécifications de notre prototype. En particulier, elle présente la plate-forme de développement GéOxygène[60] du Laboratoire COGIT de l'IGN sur laquelle nous avons mis en œuvre nos contributions, l'architecture globale du prototype, la mise en œuvre des différents modules du prototype avec les méthodes proposées.

La section 2 présente l'initialisation d'un catalogue d'éléments de vocabulaire à partir des tags OpenStreetMap suivie d'un test de notre méthode proposé à des chercheurs en géomatique du Laboratoire COGIT, afin d'illustrer comment fonctionne l'aide à la construction à la volée d'un vocabulaire.

La section 3 présente la mise en œuvre d'une méthode de correction d'incohérences, un aspect clé de notre stratégie d'intégration de contributions dans un contenu géographique collaboratif. Nous avons choisi un cas basé sur des données OpenStreetMap.

## 1 Le prototype

Cette section présente la plate-forme de développement GéOxygène et décrit les modules pertinents pour notre travail. Ensuite, nous présentons l'architecture globale de notre prototype et ses trois modules principaux.

### 1.1 La plate-forme de développement GéOxygène

GéOxygène est une plate-forme *open source* développée en Java implémentant les normes OGC/ISO pour le développement et le déploiement d'applications de recherche en SIG (Grosso et al. 2012). La Figure 67 illustre l'architecture de GéOxygène.

[60] http://oxygene-project.sourceforge.net/

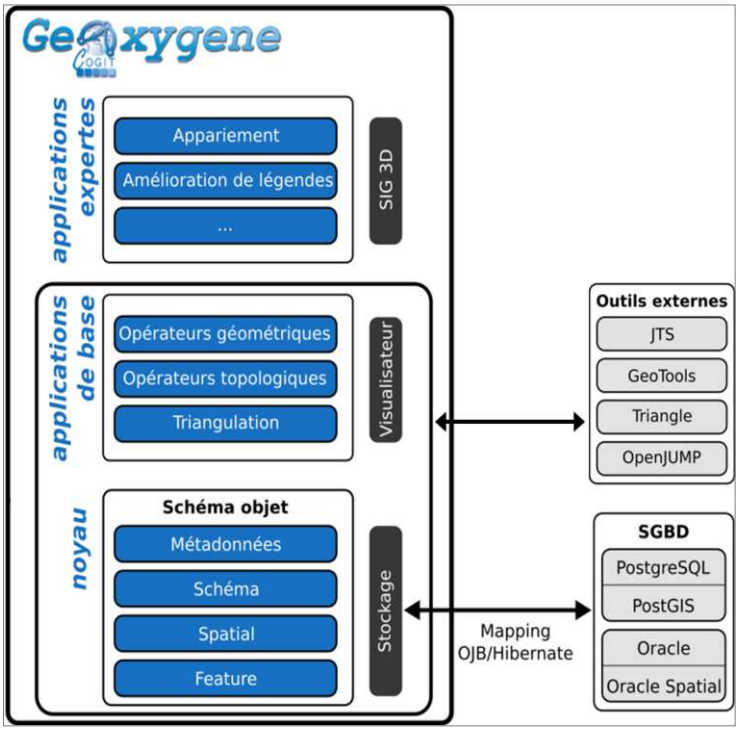



**Figure 67 : architecture de la plate-forme GéOxygène (Grosso et al. 2012)**

La plate-forme GéOxygène est composée du *noyau* qui définit les structures de données principales et d'un module contenant les *applications de base* pour la manipulation de données géographiques. GéOxygène est également composée d'un module contenant les *applications expertes* issues des travaux de recherche du COGIT. Dans ce module, il existe actuellement un sous-module de sémiologie pour l'*amélioration de légendes*, un sous-module pour la *manipulation et visualisation de données géographiques 3D*, et un sous-module d'intégration pour l'*appariement de données géographiques*. GéOxygène fournit une interface graphique qui permet la visualisation et l'édition des jeux de données géographiques ainsi que leurs schémas. De plus, GéOxygène a été récemment intégré avec la plate-forme de recherche en généralisation cartographique CartAGen (Renard et al. 2011), développée au COGIT, qui offre notamment une bibliothèque dédiée à l'analyse spatiale. Le principal avantage du choix de GéOxygène est la possibilité de réutiliser les structures ISO de données géographiques et de développer des nouvelles applications expertes pour les intégrer sous la forme de *plug-ins*.

Le *visualisateur* de GéOxygène a servi comme interface graphique de base pour le développement de l'outil client de notre prototype. Ainsi, le client a été ajouté comme un *plug-in*. Le *noyau* et le module *applications de base* ont été essentiels pour la mise en œuvre de notre modèle *Coalla* et des fonctionnalités importantes du serveur.

Du *noyau* GéOxygène, nous utilisons particulièrement des classes appartenant aux sous-modules *Schéma* et *Feature*. Plus précisément, la classe *Feature* de notre modèle *Coalla* (pour rappel voir section III.1) correspond à la classe du même nom dans le *noyau,* et représente

l'instance d'un objet géographique. La classe *GéoDonnée* de *Coalla* étend la classe *DataSet* du *noyau* et représente l'instance d'un jeu de données (un ensemble de *features*). La classe *Vocabulary* de *Coalla* étend la classe *ISOConceptualSchema* du noyau. Les classes *Feature Type*, *Relation Type* et *Property Type* de *Coalla* étendent respectivement les classes *Feature Type*, *AssociationType*, et *AttributeType* du *noyau*. Par ailleurs, nous avons décidé d'implémenter la classe *Contrainte* en nous inspirant du modèle de généralisation cartographique de CartAGen. Néanmoins, notre classe *Contrainte* de *Coalla* est une version simplifiée de la représentation en généralisation afin de ne pas ajouter une complexité inutile dans un contexte d'édition collaborative.

Nous nous servons également de plusieurs sous-modules du module *applications de base* de GéOxygène de même que des méthodes disponibles dans CartAGen pour les méthodes de correction d'incohérences. En particulier, nous utilisons les méthodes pour la création et la manipulation de *cartes topologiques* comme certains opérateurs comme l'*angle d'orientation*, la *surface minimale d'intersection*, la *translation*, la *rotation* et le *buffer* afin d'analyser les incohérences et calculer les nouvelles géométries associées à des corrections.

## 1.2 L'architecture du prototype pour l'édition collaborative

L'édition collaborative pose la question du choix d'une architecture distribuée. Nous avons choisi une architecture distribuée et centralisée comme celle du modèle client-serveur afin de nous concentrer sur les échanges entre les clients et le serveur, sur la gestion par le serveur d'un vocabulaire centralisé, et sur les stratégies du serveur pour l'évaluation et la réconciliation de contributions. La possibilité de choisir une architecture distribuée et décentralisée comme le modèle pair-à-pair (P2P) a été analysé afin de permettre le travail d'édition en mode « déconnecté » et décentralisé. Néanmoins, ce choix ajoute une complexité importante concernant la synchronisation des copies locales du contenu entre les clients P2P. Nous avons donc choisi une approche centralisée client-serveur. Nous avons conçu un client de type « lourd » afin de faciliter la reprise du *visualisateur* de GéOxygène du côté client. L'architecture générale du prototype pour l'édition collaborative d'un contenu géographique et la gestion de sa cohérence est illustrée en Figure 68.

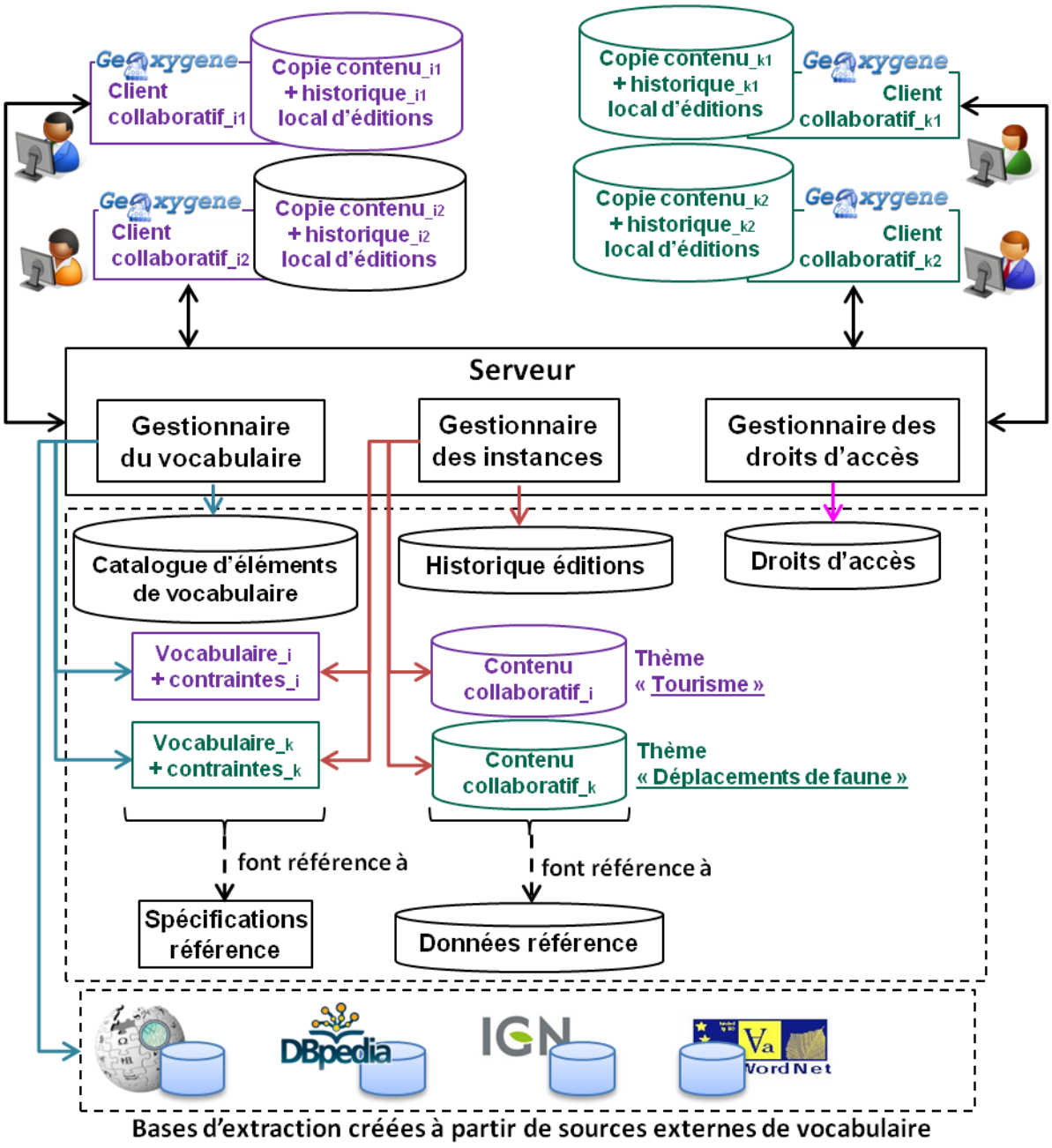



**Figure 68 : diagramme d'architecture de notre prototype pour l'édition collaborative d'un contenu géographique et la gestion de sa cohérence**

### 1.2.1 Le client collaboratif

Un *module d'édition collaborative* a été ajouté dans le *visualisateur* de GéOxygène comme un *plug-in* afin de construire un *client collaboratif* « lourd ». La Figure 69 montre l'interface graphique du client collaboratif.

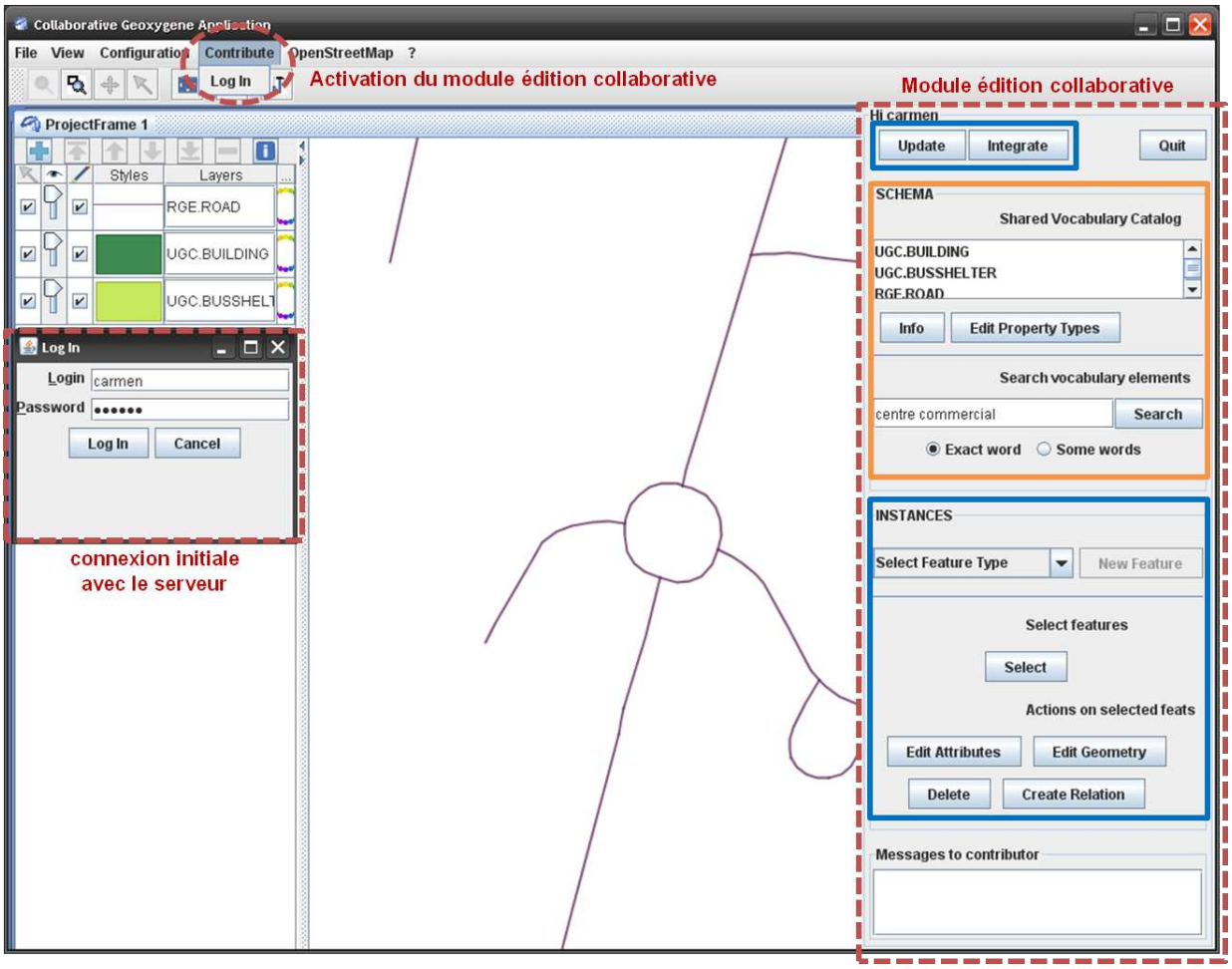



**Figure 69 : Un plug-in a été ajouté dans le visualisateur de GéOxygène afin de construire ce client collaboratif « lourd »**

Le module d'édition collaborative du client est activé via le menu de connexion *login* (voir panel et menu entourés en pointillée rouge en Figure 69). Le contributeur s'est identifié sur le serveur et reçoit un *identifiant de la session* qui est valide pendant la durée de sa session. Il télécharge ensuite depuis le serveur (voir bouton *update* dans le cadre bleu en Figure 69), les dernières versions des contenus auxquels il a un droit d'accès. Pour modifier ces contenus, il doit en effet avoir un droit d'accès en écriture. De plus, il peut avoir un droit d'accès en lecture sur des contenus issus d'autres contributeurs. Pendant que la session est active, le serveur conserve aussi cette copie du contenu et lui associe l'*identifiant de la session du client*. Les données téléchargées, instances de `Feature` et de `Relation`, correspondent à une zone géographique qu'il précise en dessinant une enveloppe, ainsi qu'à des `Types de Feature` qu'il sélectionne auparavant. Un `Feature` contient un identifiant (temporaire pour les nouveaux `Features` créées), une géométrie, des propriétés initialisées et un `Type de Feature` (sauf pour les `Features` du type *Thing*). Une `Relation` fait référence aux identifiants des deux `Features` qui participent à la relation et le nom du `Type de Relation` correspondant (ex : `en face de`). Ce type de relation figure parmi son vocabulaire : une instance créée par lui-même ou

copiée auparavant depuis le *catalogue d'éléments de vocabulaire*. Pendant sa session, les `Features` et les relations édités par le contributeur sont stockées dans la `Copie Locale du Contenu Central` du client. Les `Features` sont stockés comme un jeu de données (un `DataSet` GéOxygène) par `Type de Feature` et les relations sont stockées dans une liste. A partir de là, le contributeur peut éditer ses instances en effectuant les opérations d'édition disponibles sur *Coalla* (voir section III.3.1) en utilisant l'interface d'édition comme par exemple, modifier un *feature* ou créer une relation (voir panel bleu en Figure 69). Les opérations effectuées seront stockées avec des identifiants locaux dans l'`Historique Local d'Editions` du contributeur. Le contributeur peut ensuite intégrer ces contributions dans le contenu central en appuyant sur le bouton *integrate* (voir le cadre bleu en Figure 69). Par exemple, la séquence d'édition du contributeur `Cbrando` voulant créer un nouveau `Feature` de type `Abribus` (un `Type de Feature` instancié dans le vocabulaire commun) tout en assignant une géométrie polygonale[61] est stockée dans l'`Historique Local d'Editions`, puis transférée vers le serveur de la manière suivante :

```
<editions>
   <creerFeature>
      <changeContenu> oui </changeContenu>
      <idEdition> 1 </idEdition>
      <genereeParSysteme> non </genereeParSysteme>
      <timestamp> 2012-09-26T10:56:17.578+02:00 </timestamp>
      <idFeature> -2 </idFeature>
      <nomTypeFeature> Abribus </featureTypeName>
      <contributeur> CBrando </contributeur>
   </creerFeature>

   <modifierValeurProprieteFeature>
      <changeContenu> oui </changeContenu>
      <idEdition> 2 </idEdition>
      <genereeParSysteme> non </genereeParSysteme>
      <timestamp> 2012-09-25T13:40:26.309+02:00 </timestamp>
      <idFeature> -2 </idFeature>
      <nomTypeFeature> Abribus </nomTypeFeature>
      <estGeometrie> oui </estGeometrie>
      <geomChaine> POLYGON ((886425.8758432227 2048617.7123666063, … ,
      886425.8758432227 2048617.7123666063)) </geomChaine>
      <contributeur> CBrando </contributeur>
    </modifierValeurProprieteFeature>
</editions>
```

Pour modifier son vocabulaire, le contributeur peut éditer les éléments de vocabulaire correspondant via un formulaire ou initier le processus d'aide à la construction du vocabulaire afin d'avoir, à partir de mots-clés, des suggestions de nouveaux éléments pour son vocabulaire (voir panel dans le cadre orange en Figure 69).

[61] La géométrie est dans le système de coordonnées Lambert93

### 1.2.2 Le serveur

Le serveur est composé des trois modules suivants : le *gestionnaire du vocabulaire*, le *gestionnaire d'instances* et le *gestionnaire de droits d'accès*. Nous les décrivons dans les sections suivantes.

### 1.2.3 Le serveur : le module Gestionnaire du vocabulaire

Le *module gestionnaire du vocabulaire* vise à aider un contributeur à constituer et à modifier le *vocabulaire commun* aux contributeurs. Cette aide est déclenchée par le contributeur en tapant un mot-clé composé ou non (ex : `centre commercial`) et en appuyant sur le bouton *search* sur l'interface graphique du *client collaboratif* (voir panel encadré en orange en Figure 69). Ensuite, le serveur reçoit le mot-clé et lance le *processus de construction d'un vocabulaire formel* décrit dans la section III.2.

Le *module gestionnaire du vocabulaire* interroge d'abord le *catalogue d'éléments de vocabulaire* afin de trouver des éléments de vocabulaire déjà instanciés par le système et qu'il voudrait copier dans son vocabulaire. Ainsi, nous initialisons le *catalogue d'éléments de vocabulaire* afin de mettre à disposition pour les contributeurs des éléments de vocabulaire déjà instanciés (ex : un `Type de Feature autoroute`, un `Type de relation intersection`) qui peuvent être potentiellement pertinents. Certains éléments ont été auparavant extraits grâce au *processus de construction d'un vocabulaire formel* à partir d'un ensemble prédéfini de mots-clés. D'autres éléments ont été manuellement initialisés. Quand un contributeur sélectionne un de ces éléments, son instance correspondante dans le *catalogue d'éléments de vocabulaire* est modifiée afin d'indiquer qu'un contributeur particulier l'a utilisé. En proposant des suggestions au contributeur, le *gestionnaire du vocabulaire* encourage la réutilisation des éléments dans ce catalogue qui ont été auparavant jugés pertinentes et instanciés par d'autres contributeurs.

En revanche, si aucun élément de vocabulaire n'a été trouvé dans le *catalogue d'éléments de vocabulaire* à partir du mot-clé indiqué, le *processus de construction d'un vocabulaire formel* interroge à la volée les quatre bases que nous avons constituées à partir de sources externes de vocabulaire : la base d'extraction Wikipédia, DBpedia, WordNet, et IGN.

Ci-dessous, nous décrivons les deux cas de recherche d'éléments de vocabulaire signalés auparavant : la recherche est effectuée ou bien dans le *catalogue d'éléments de vocabulaire* ou alors dans les bases d'extraction. Dans les deux cas, nous montrons un exemple d'éléments de vocabulaire extrait à partir du mot-clé `centre commercial`. Ensuite, nous expliquons la manière dont des éléments sélectionnés par le contributeur sont instanciés dans le *vocabulaire commun* et le *catalogue d'éléments de vocabulaire*.

**Cas : recherche à partir du catalogue d'éléments de vocabulaire**

Ici, nous détaillons le *catalogue d'éléments de vocabulaire* et la manière dont les éléments trouvés à partir du mot-clé, sont présentés au contributeur sur son interface du *client collaboratif* et ensuite intégrés, s'il les a sélectionnés, dans le *vocabulaire commun*.

Le *catalogue d'éléments de vocabulaire* contient des instances d'éléments de vocabulaire : `Type de Feature`, `Type de Relation` et `Type de Propriété`. Ces instances sont représentées en XML et stockées dans une base de données native XML EXist[62] (Meier 2003). Par exemple, prenons le cas d'un `Type de Feature` appelé `centre commercial` extrait grâce au *processus de construction d'un vocabulaire formel*. Il a été extrait auparavant par notre processus à partir du graphe de catégories de Wikipédia (WCG), c'est-à-dire qu'il correspond à une catégorie du WCG. Puis, ce `Type de Feature centre commercial` a été instancié par deux contributeurs prénommés `Angela` et `Félix`.

```
<featureType>
      <nom> Centre commercial </nom>
      <instanciePar> Angela </instanciePar>
      <instanciePar> Félix </instanciePar>
      <source>
            <nom> graphe de catégories de Wikipédia (WCG)</nom>
            <categorie>
                  <nom> Centre commercial </nom>
                  <superCatWCG> Bâtiment et local de commerce </superCatWCG>
                  <superCatWCG> Entreprise de distribution </superCatWCG>
            </categorie>
      </source>
</featureType>
```

Nous avons également instancié dans le *catalogue d'éléments de vocabulaire* des `Types de Relation` pertinents pour la gestion de la cohérence définis entre des `Types de Features` du contenu collaboratif et du contenu de référence. Ces `Types de Relations` servent au contributeur à signaler, au niveau du modèle, les relations qui doivent être préservées entre le contenu collaboratif et le contenu de référence. Prenons le cas d'un `Type de Relation` appelé `intersection` que nous avons prédéfini manuellement dans le *catalogue d'éléments de vocabulaire*, entre un `Type de feature bâtiment` et un `Type de Feature route` provenant des spécifications IGN. Il a déjà été instancié par le contributeur `Cbrando`. Cet élément a été stocké en XML dans le *catalogue d'éléments* de la manière suivante :

```
<relationType>
      <nom> intersection </nom>
      <instanciePar> CBrando </instanciePar>
      <topologique> oui </topologique>
      <featureType1>
```

[62] http://www.exist-db.org

```
            <nom> bâtiment </nom>
      </featureType1>
      <featureType2>
            <nom> route </nom>
            <estIGN> oui </estIGN>
      <featureType2>
</relationType>
```

Cette instance de `Type de Relation` peut être utilisée pour définir une contrainte d'intégrité si le contributeur le souhaite, par exemple, « les bâtiments ne chevauchent pas les routes de l'IGN ». Ainsi, la recherche à partir du mot-clé `centre commercial` renvoie le `Type de Feature` et le `Type de Relation` ci-dessous tous deux extraits du *catalogue d'éléments de vocabulaire*. Ensuite, ce résultat est transféré vers le client dans une liste d'*éléments partiels de vocabulaire* qui est présentée au contributeur dans une nouvelle fenêtre (voir la Figure 70). Il sélectionne les éléments de la liste qui l'intéressent, en effaçant de la zone de texte ceux qui ne l'intéressent pas. Ensuite, il appuie sur le bouton *save* (voir la Figure 70) pour enregistrer dans le *vocabulaire commun*, les éléments qui ont été pertinents pour lui. Le client transfère ensuite vers le serveur l'identifiant local des éléments d'intérêt. Les instances des éléments sélectionnés sont mises à jour dans le catalogue, comme la valeur du `instanciePar`. Des copies des éléments sélectionnés sont créées dans le *vocabulaire commun*.

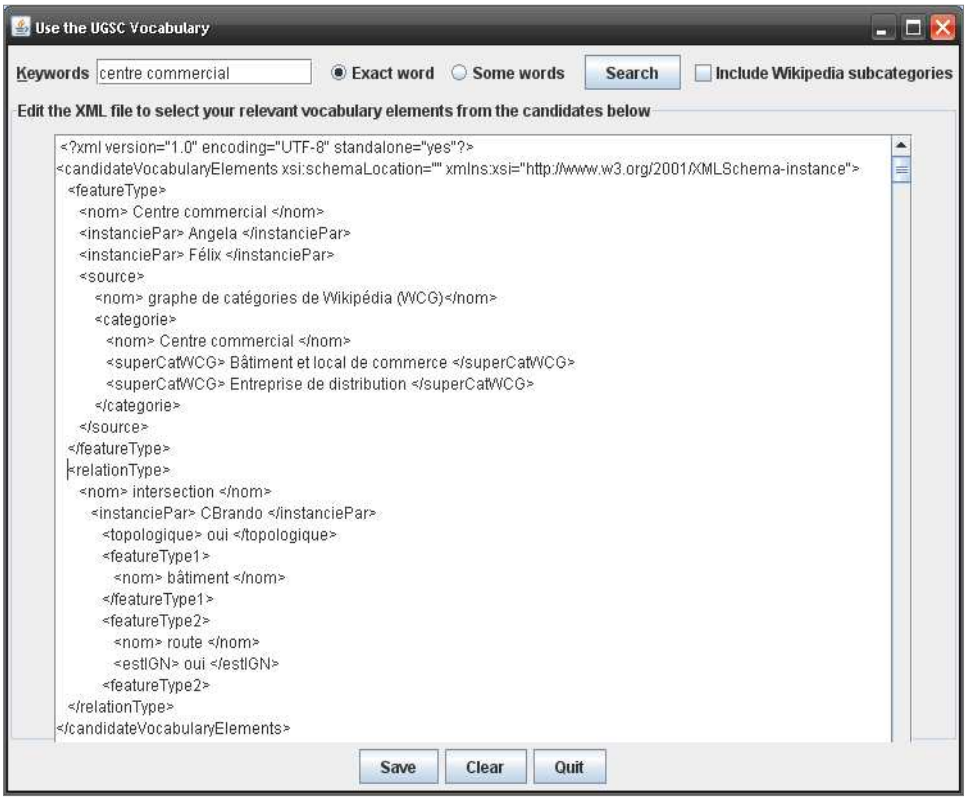



**Figure 70 : éléments de vocabulaire trouvés dans le catalogue d'éléments de vocabulaire par le processus de construction de vocabulaire à partir du mot-clé `centre commercial`**

Nous sommes bien conscients que l'interface graphique développée n'est pas très ergonomique car les résultats de la méthode s'affichent en XML. En effet, ce point-là est très important pour améliorer l'expérience pour le contributeur. Néanmoins, nous ne nous sommes pas concentrés sur l'interface graphique mais sur la fonctionnalité.

**Cas : recherche à partir des bases d'extraction**

Ici, nous décrivons les *bases d'extraction* et la manière dont les éléments de vocabulaire sont extraits (voir la Figure 71). Puis, nous expliquons la façon dont les éléments trouvés à partir du mot-clé sont présentés au contributeur sur son interface du *client collaboratif*. Nous détaillons aussi la façon dont les éléments sélectionnés par le contributeur sont ensuite intégrés dans le *vocabulaire commun* de même que dans le *catalogue d'éléments de vocabulaire* pour qu'ils puissent être réutilisés par d'autres contributeurs.

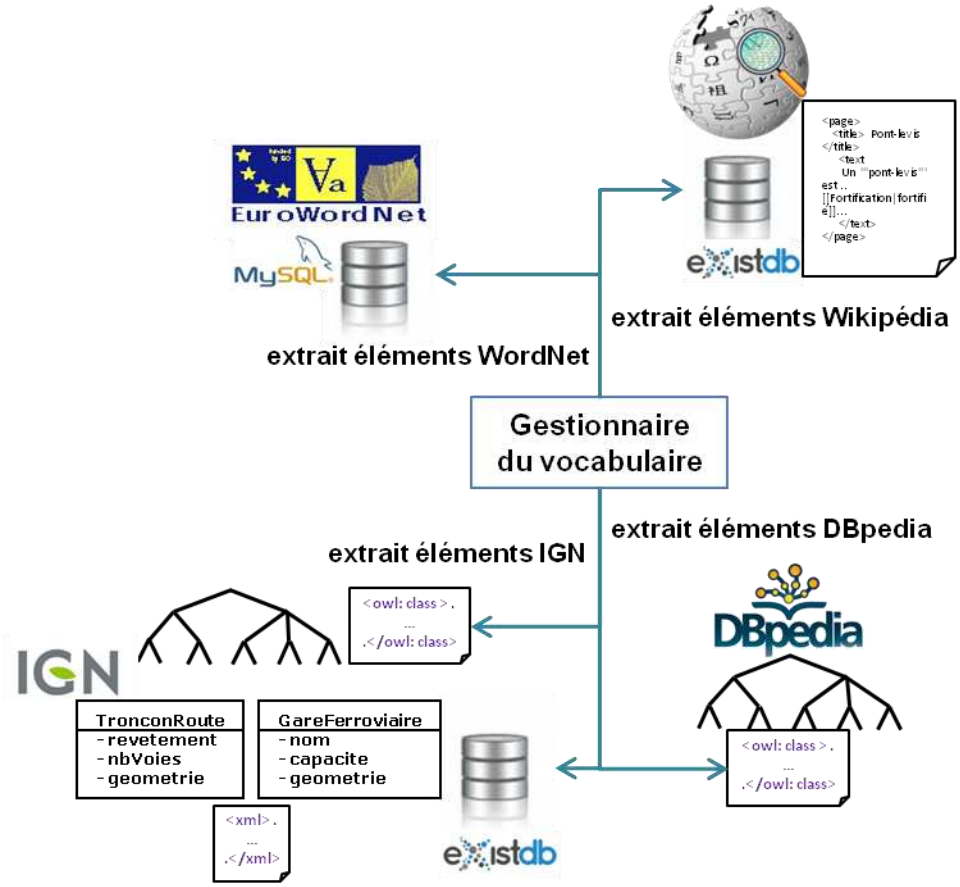



**Figure 71 : diagramme d'extraction des éléments de vocabulaire : Wikipédia, DBpedia, WordNet, et IGN à partir des sources externes**

La base d'extraction Wikipédia

Cette base contient la version francophone datée du 13 mars 2012 de Wikipédia. Ce contenu est périodiquement disponible sur le site de téléchargement de la fondation Wikimédia[63] sous la forme d'un fichier « dump » XML de 8.8 Go contenant 1.229.970 pages. Ce contenu est stocké dans une base de données native XML EXist et est interrogé via son API Java et d'une manière efficace grâce à des index natifs dédiés au texte.

Le *processus de construction d'un vocabulaire formel* extrait, à partir du mot-clé, les `éléments Wikipédia` suivants : les `Catégories` et les `Modèles Infoboxes` (voir section III.2.1). Deux `éléments Wikipédia` résultant de la recherche par le mot-clé `centre commercial` dans la *base d'extraction Wikipédia* sont montrés en XML ci-dessous :

```
<categorie>
      <nom> Centre commercial </nom>
      <superCatWCG> Bâtiment et local de commerce </superCatWCG>
      <superCatWCG> Entreprise de distribution </superCatWCG>
</categorie>

<modeleInfobox>
      <nom> Centre commercial </nom>
      <champ> nom </champ>
      <champ> pays </champ>
      <champ> propriétaire </champ>
      <champ> date d'ouverture </champ>
      <champ> latitude </champ>
      …
</modeleInfobox>
```

La base d'extraction DBpedia

Cette base correspond à l'ontologie DBpedia formalisée en OWL en version 3.8 datée d'août 2012. Ce contenu est interrogé via l'API Java Jena[64].

Le *processus de construction d'un vocabulaire formel* extrait à la volée, à partir du mot-clé, les `éléments DBpedia` suivants : les `Types de Propriétés Onto` et les `Types Relations Onto` (voir la section III.2.2). Deux éléments DBpedia, un *Type de Propriété Onto* nombre `d'étages` et un *Type de Relation* Onto `bâtiment significatif`, résultant de la recherche par le mot-clé `centre commercial` dans la *base d'extraction DBpedia* sont montrés en XML ci-dessous :

```
<typeProprieteOnto>
      <nom> nombre d'étages </nom>
```

[63] http://dumps.wikimedia.org/frwiki/
[64] http://jena.apache.org/

```
      <domaine> bâtiment </domaine>
      <range> entier non négatif </range>
</typeProprieteOnto>

<typeRelationOnto>
      <nom> bâtiment significatif </nom>
      <domaine> bâtiment </domaine>
      <range> architecte </range>
</typeRelationOnto>
```

La base d'extraction WordNet

Cette base est une version française de la base de données lexicale WordNet en version 1.5. Elle a été constituée sous la forme d'une base relationnelle de données MySQL[65] qui contient trois tables relationnelles principales correspondant respectivement à des littéraux, des méronymies entre ces littéraux, et des hyperonymies entre ces littéraux. Elle est interrogée grâce à un connecteur Java JDBC.

Le *processus de construction d'un vocabulaire formel* extrait, à partir du mot-clé, les `éléments WordNet` suivants : les `Types de Relation « Méronymie »` les `Types de Relation « Hyperonymie »` (voir la section III.2.3). Un élément WordNet résultant de la recherche par le mot-clé `centre commercial` dans la *base d'extraction WordNet* est montré en XML ci-dessous :

```
<typeRelationWordNet>
      <type> Hyperonymie </type>
      <literal1> bâtiment </literal1>
      <literal2> hall </literal2>
</typeRelationOntoWordNet>
```

La base d'extraction IGN

Cette base correspond à l'ontologie IGN formalisée en OWL et interrogée via l'API Java Jena. Elle correspond de plus à une version formelle en XML de la description du schéma produit de la BDTopo® de l'IGN, stockée dans une base de données XML EXist[66] et interrogée via l'API Java EXist.

Le *processus de construction d'un vocabulaire formel* extrait, à partir du mot-clé, les `éléments IGN` suivants : les `Concepts` (dans l'ontologie) et les `Classes` (du schéma produit) (voir la section III.2.4). Un élément IGN résultant de la recherche par le mot-clé `centre commercial` dans la *base d'extraction IGN* est montré en XML ci-dessous :

```
<classe>
```

[65] http://dev.mysql.com/downloads/mysql/
[66] http://www.exist-db.org

```
      <nom>Point d'activité ou d'intérêt</nom>
      <description>
            Objet ponctuel localisant un équipement public, un
            site ou une zone ayant un caractère administratif, culturel,
            sportif, industriel ou commercial.
      </description>
      <valeurAttributNature>Divers commercial</valuesForNatureAttribute>
</classe>
```

**Instanciation des éléments sélectionnés du vocabulaire dans le vocabulaire commun**

Des `éléments Wikipédia, DBpedia, WordNet et IGN`, comme ceux extraits ci-dessus, sont transférés dans une liste vers le client et présentés au contributeur via l'interface graphique de la Figure 70. Le contributeur sélectionne les éléments qui l'intéressent, en effaçant de la zone de texte ceux qui ne l'intéressent pas. Ensuite, il appuie sur le bouton *save* (voir la Figure 70) pour enregistrer ces éléments dans le *vocabulaire commun* et dans le *catalogue d'éléments de vocabulaire* pour pouvoir être éventuellement réutilisés par d'autres contributeurs. Ainsi, le client transfère vers le serveur l'identifiant local des éléments choisis.

Le *module gestionnaire du vocabulaire* initialise des instances de `Type de Feature`, de `Type de Propriété` et de `Type de Relation` dans le *vocabulaire commun* selon les éléments sources (`éléments Wikipédia, DBpedia, WordNet et IGN`), comme expliqué dans la section III.2, de même que dans le *catalogue d'éléments de vocabulaire* pour une réutilisation par d'autres contributeurs.

Par exemple, une instance d'un `Type de Propriété` est initialisée dans le *vocabulaire commun* et le *catalogue d'éléments de vocabulaire* à partir du champ `propriétaire` du `Modèle Infobox centre commercial` sélectionné par le contributeur `CBrando`. De manière similaire, une instance d'un `Type de Feature` est initialisée à partir de la `Catégorie centre commercial` du WCG. Ces éléments sont représentés ainsi :

```
<featureType>
      <nom> Centre commercial </nom>
      <instanciePar> Cbrando </instanciePar>
      <source>
            <nom> graphe de catégories de Wikipédia (WCG)</nom>
            <categorie>
                  <nom> Centre commercial </nom>
                  <superCatWCG> Bâtiment et local de commerce </superCatWCG>
                  <superCatWCG> Entreprise de distribution </superCatWCG>
            </categorie>
      </source>
</featureType>

<proprieteType>
      <nom> propriétaire </nom>
      <nomFeatureType> centre commercial </nomFeatureType>
```

```
      <instanciePar> Cbrando </instanciePar>
      <source>
            <nom> Modèles Infobox Wikipédia </nom>
            <modeleInfobox>
                  <nom> Centre commercial </nom>
                  <champ> propriétaire </champ>
            </modeleInfobox>
      </source>
</proprieteType>
```

Le *gestionnaire du vocabulaire* envoie une confirmation au client indiquant que le *vocabulaire commun* a été mis à jour. Le vocabulaire de la `Copie Locale du Contenu` du client est également mis à jour. Ainsi, par exemple, des instances de *features* correspondant au `Type de Feature centre commercial` pourront être créés avec une nouvelle propriété `propriétaire` à remplir par le contributeur, tout en conformité avec le *vocabulaire commun*.

### 1.2.4 Le serveur : le module Gestionnaire d'instances

Le *module gestionnaire d'instances* vise à aider un contributeur à intégrer ses contributions saisies sur sa *copie locale et partielle du contenu* dans le *contenu collaboratif*. Une fois que le contributeur a terminé d'éditer ses données via les fonctionnalités d'édition disponibles sur le *client collaboratif* (voir le deuxième panel encadré en bleu en Figure 72), il déclenche cette aide en appuyant sur le bouton *integrate* sur l'interface graphique du *client collaboratif* (voir le premier panel encadré en bleu en Figure 72). La Figure 72 présente l'exemple d'un contributeur qui saisit un nouvel abribus chevauchant un tronçon de route appartenant au contenu de référence IGN.

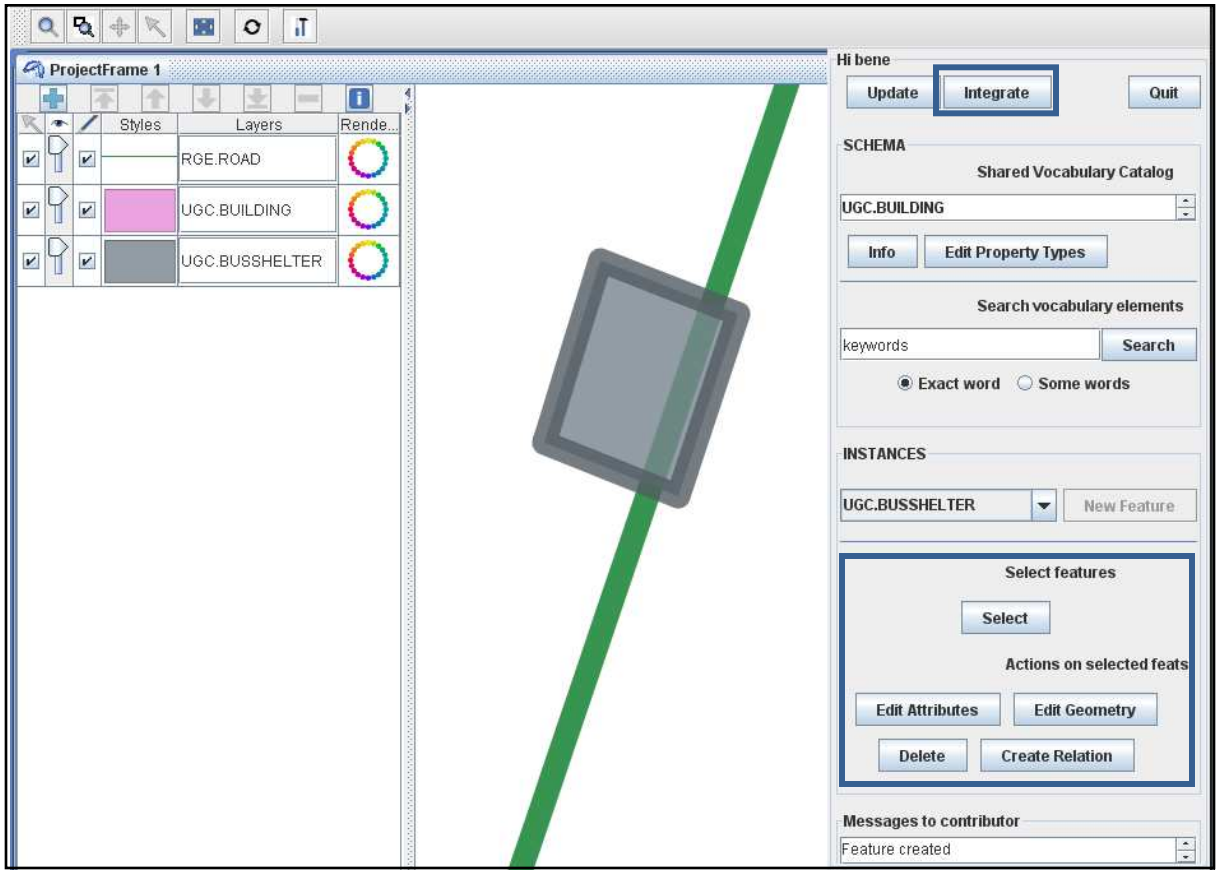



**Figure 72 : saisie d'un abribus par un contributeur sur l'interface graphique du client collaboratif**

Ces contributions sont transférées vers le serveur sous la forme d'une séquence d'éditions. Ces éditions sont stockées ensuite dans l'`Historique d'Editions`. Le code ci-dessous est une séquence d'éditions concernant la création du `Feature`, c'est-à-dire l'initialisation de son identifiant et son type, puis la modification du `Feature` pour assigner une géométrie :

```
<editions>
   <creerFeature>
      <changeContenu> oui </changeContenu>
      <idEdition> 1 </idEdition>
      <genereeParSysteme> non </genereeParSysteme>
      <timestamp> 2012-09-26T10:56:17.578+02:00 </timestamp>
      <idFeature> -2 </idFeature>
      <nomTypeFeature> Abribus </featureTypeName>
      <contributeur> CBrando </contributeur>
   </creerFeature>

   <modifierValeurProprieteFeature>
      <changeContenu> oui </changeContenu>
      <idEdition> 2 </idEdition>
      <genereeParSysteme> non </genereeParSysteme>
```

```
      <timestamp> 2012-09-25T13:40:26.309+02:00 </timestamp>
      <idFeature> -2 </idFeature>
      <nomTypeFeature> Abribus </nomTypeFeature>
      <estGeometrie> oui </estGeometrie>
      <geomChaine> POLYGON ((886425.8758432227 2048617.7123666063, … ,
      886425.8758432227 2048617.7123666063)) </geomChaine>
      <contributeur> CBrando </contributeur>
    </modifierValeurProprieteFeature>
</editions>
```

Le *gestionnaire des instances* possède une *file d'attente* dans l'ordre chronologique (First-In-First-Out) afin de traiter les séquences d'éditions qui arrivent des clients connectés. En présence de plus d'une séquence à évaluer dans la *fille d'attente*, le *gestionnaire des instances* lance la *stratégie de réconciliation de deux séquences d'éditions*, décrite en section III.3.2.2, pour chaque paire de séquences. Ensuite, le *gestionnaire des instances* lance la *stratégie d'évaluation d'une séquence d'éditions*, décrite en section III.3.2.1, pour les séquences fusionnées et pour le cas d'une seule séquence d'éditions dans la file d'attente.

Ci-dessous, nous allons décrire la façon dont nous avons implémenté certains éléments clés utilisés par les deux stratégies : la base de contraintes, les jeux de données d'évaluation, les jeux de données temporaires, et la présentation des corrections sur le client collaboratif.

**Implémentation de la base de contraintes**

Ces contraintes peuvent faire référence à des spécifications de données de référence afin de gérer la cohérence (voir section III.1.3). Par exemple, la contrainte `les abribus ne chevauchent pas les routes de l'IGN` est définie à partir du *Type de Relation* `intersection` entre les abribus du contenu collaboratifs et les tronçons de routes du contenu de référence, instanciés dans le *vocabulaire commun*. Elle est représentée en XML dans la base de contraintes ainsi :

```
<contrainteTypeRelation>
      <label> les abribus ne chevauchent pas les routes de l'IGN </label>
      <messageViolation> l'abribus proposé chevauche le tronçon de route de
      référence </messageViolation>
      <estInterdiction> oui </estInterdiction>
      <relationType>
            <nom> intersection </nom>
            <estSpatiale> oui </estSpatiale>
            <featureType1>
                  <nom> abribus </nom>
            </featureType1>
            <featureType2>
                  <nom> route </nom>
                  <estIGN> oui </estIGN>
            <featureType2>
      </relationType>
```

```
</contrainteTypeRelation>
```

Les contraintes dans la base peuvent indiquer aussi une dépendance entre des `Types de Propriétés` (voir section III.1.4). Par exemple, la contrainte de dépendance entre les `Type de Propriétés` : `aire` et `géométrie` est représentée en XML ainsi :

```
<contrainteDepedenceTypePropriete>
      <label> depedence entre longueur et géométrie d'un abribus </label>
      <messageViolation> la propriété longueur d'un abribus dépend de sa
      géométrie </messageViolation>
      <proprieteType>
            <nom> longueur </nom>
            <estSpatiale> non </estSpatiale>
      </proprieteType>
      <proprieteType>
            <nom> geometrie </nom>
            <estSpatiale> oui </estSpatiale>
      </proprieteType>
</contrainteDepedenceTypePropriete>
```

Représenter les contraintes en XML est un choix qui a pour but de pouvoir les échanger entre les clients et le serveur. Ce choix permet aussi aux contributeurs de définir d'autres contraintes facilement en modifiant le code XML. A partir de cette base de contraintes, la *stratégie d'évaluation d'une séquence d'éditions* peut consulter et récupérer les contraintes potentiellement violées en fonction du `Type de Feature` (ou `Type de relation`) des `Features` (ou des `Relations`) touchés par la séquence d'éditions.

**Implémentation des jeux de données d'évaluation et des jeux de données temporaires**

Ces jeux de données sont créés périodiquement par la *stratégie d'évaluation d'une séquence d'éditions* afin d'évaluer les contraintes potentiellement violées. Ces jeux de données représentent des instances de jeux de données implémentées comme des `DataSets` dans GéOxygène. Un jeu de données est un ensemble de `Features` d'un même `Type de Feature` et chacun avec des identifiants. Pour extraire un jeu de données d'évaluation et pour créer un jeu de données temporaire, la *stratégie d'évaluation d'une séquence d'éditions* crée des instances correspondantes de `DataSets`. De plus, la stratégie se sert de la `Copie Locale du Contenu` du client, conservée du côté serveur, afin d'effectuer les extractions des jeux de données et d'éviter des échanges inutiles avec le client.

**Présentation des corrections sur le client collaboratif**

Le *gestionnaire des instances* prépare sa réponse en XML pour le client en explicitant l'information sur la contrainte violée et une séquence d'éditions correctrices, comme par exemple, ci-dessous :

```
<corrections>
   <correction>
      <incoherenceTypeRelation>
         <contrainteTypeRelationViolee>
            <contrainteTypeRelation>
            <label> les abribus ne chevauchent pas les routes de l'IGN
            </label>
            …
         </contrainteTypeRelationViolee>
      </incoherenceTypeRelation>

      <editionsCorrectrices>
            <modifierValeurProprieteFeature>
                  <changeContenu> oui </changeContenu>
                  <editionId> 3 </editionId>
                  <estGenereeParSysteme> true </estGenereeParSysteme>
                  <timestamp> 2012-09-26T10:56:20.877+02:00 </timestamp>
                  <idFeature> 1 </idFeature>
                  <featureTypeName> Abribus </featureTypeName>
                  <concerneGeometrie> oui </concerneGeometrie>
                  <geomString> POLYGON ((886413.1277131666 2048627.624298337,
                  …, 886413.1277131666 2048627.624298337)) </geomString>
            </modifierValeurProprieteFeature>
      </editionsCorrectrices>

      <correctionEffectuee> shift </correctionEffectuee>
   </correction>
</corrections>
```

Du point de vue du contributeur, la Figure 73 montre la manière dont le contributeur visualise la proposition du serveur de modifier la géométrie de l'abribus qui ne cause pas une incohérence et qui reste « proche » de la géométrie saisie par le contributeur.

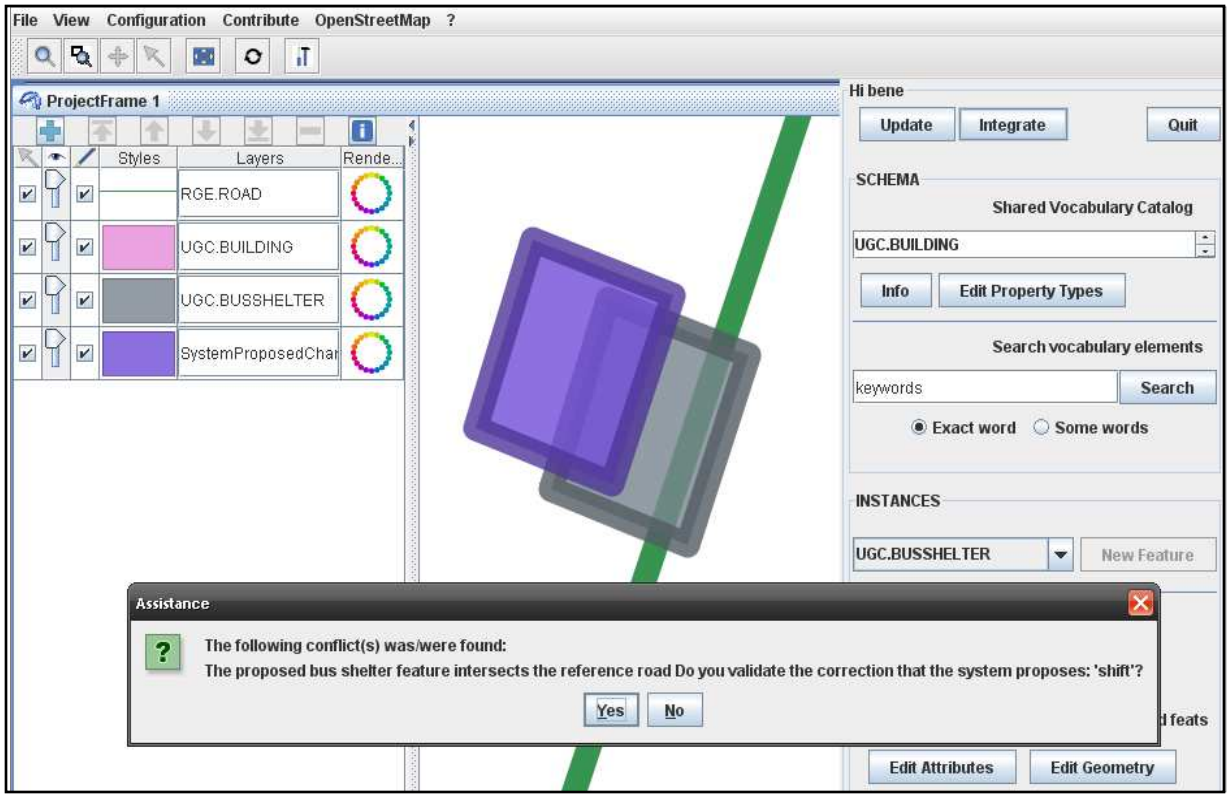



**Figure 73 : réponse du serveur au client après la soumission d'une séquence d'éditions**

### 1.2.5 Le serveur : le module Gestionnaire des droits d'accès

Au moment de la connexion du client sur le serveur, le *module gestionnaire des droits d'accès* vérifie si un contributeur peut accéder un mode `lecture`, `modification`, ou `suppression`, au contenu collaboratif en fournissant sons identifiant et son mot de passe. La granularité des droits d'accès au contenu est définie au niveau d'une instance (objet ou relation) ou des instances d'une classe (si objet) ou d'un type de relation (si relation).

# 2 Aider à la construction d'un vocabulaire formel

Cette section présente l'initialisation d'un *catalogue d'éléments de vocabulaire* à partir de notre version française de l'ontologie OSMonto (Codescu et al. 2011) qui a été construite à partir des tags OSM et décrite en section II.2.1. Cette version française est utilisée comme source d'entrée de vocabulaire par notre *processus de construction d'un vocabulaire formel*. Ensuite, nous introduisons la notion de *taux de correspondance* que nous utilisons pour comparer la version française d'OSMonto avec les sources externes de vocabulaire, afin d'établir la couverture par thèmes d'un vocabulaire par rapport à un autre. Nous utilisons la *méthode*

*d'extraction de types de relations à partir des articles Wikipédia*, afin d'enrichir ce catalogue d'éléments de vocabulaire avec des types de relations, en particulier sur le thème de villes.

Une fois que notre catalogue est initialisé, nous effectuons le test du *processus de construction d'un vocabulaire formel* sur des chercheurs en géomatique du Laboratoire COGIT qui voudraient construire un contenu collaboratif ciblé sur leurs projets de recherche dans des thématiques diverses (ex : transport en commun, tourisme).

## 2.1 Initialisation d'un catalogue d'éléments de vocabulaire à partir d'une version française d'OSMonto

D'abord, nous expliquons la manière dont nous avons construit notre version française d'OSMOnto. Ensuite, nous décrivons les différentes étapes d'extraction de notre *processus de construction d'un vocabulaire formel* qui utilise comme source d'entrée cette version française d'OSMonto. Nous utilisons la notion de *taux de correspondance* pour comparer la source entrée avec les différentes sources externes de vocabulaire utilisées par le processus : Wikipédia, DBpedia, WordNet, et IGN. Nous décrivons également des éléments de vocabulaire construits qui sont conservés dans notre *catalogue d'éléments de vocabulaire*.

### 2.1.1 Construction d'une version française d'OSMonto : OSMonto-fr

D'abord, il faut rappeler que la profondeur arborescente d'OSMonto est égale à un, les classes niveau #2 correspondent à des thèmes (ex : commerces) et les classes niveau #1 à leurs classes (ex : boulangerie) (voir la section II.2.1). La première colonne du Tableau 9 liste les 19 thèmes (ou classes niveau #2) d'OSMonto (avec ses labels en français) choisis à partir des 21 classes niveau #2 d'OSMonto. Particulièrement, les classes `k_cuisine` et `k_emergency` ont été exclues dans notre version. La deuxième colonne du Tableau 9 montre, par thème, le nombre des classes OSMonto niveau #1 dont au moins un label en français a été ajouté. Plus précisément, un à trois labels en français ont été manuellement ajoutés à chacune de ces classes en se servant de la version française de la documentation du site Map Features[67] ou, en absence d'une traduction des dictionnaires en ligne sont utilisés. La liste entière des 257 classes niveau #1 de l'ontologie OSMonto avec leurs labels en français est présentée dans l'Annexe B.

| Thèmes | Nombre de classes correspondantes |
|---|---|
| k_amenity (Equipements publics) | 73 |
| k_shop (Commerces) | 45 |
| k_natural (Formation végétale) | 21 |
| k_leisure (Loisirs) | 16 |
| k_tourism (Tourisme) | 15 |
| k_highway (Routes) | 13 |

[67] http://wiki.openstreetmap.org/wiki/FR:Map_Features

| | |
|---|---|
| k_waterway (Cours d'eau) | 9 |
| k_railway (Chemins de fer) | 9 |
| k_man_made (Edifices) | 8 |
| k_route (Itinéraires) | 8 |
| k_historic (Patrimoine) | 7 |
| k_power (Energie) | 7 |
| k_military (Défense) | 6 |
| k_aeroway (Aviation) | 6 |
| k_boundary (Frontières) | 4 |
| k_aerialway (Transports par câble) | 4 |
| k_power_source (Centrales d'électricité) | 4 |
| k_landuse (Utilisation du sol) | 1 |
| k_station (Stations de transport) | 1 |
| **Total** | 257 |

**Tableau 9 : les 19 thèmes (ou classes niveau #2) d'OSMonto et par thème le nombre des classes OSMonto niveau #1 correspondantes sur lesquelles au moins un label en français a été ajouté**

La Figure 74 (à g.) montre un exemple des classes que nous avons conservé pour OSMonto-fr, particulièrement, la classe OSMonto niveau #2 et ses 4 classes niveau#1 correspondantes : `v_protected_area` (zone protégée), `v_national_park` (parc national), `v_museum` (limite administrative) et `v_maritime` (frontière maritime). En revanche, la Figure 74 (à d.) montre un exemple des classes que nous avons exclues d' OSMonto-fr, en particulier, les classes niveau #2 d'OSMonto nommées *`can_have …`* (il peut avoir …). Ce choix d'exclusion a pour but d'éviter les doublons de certaines classes. Par exemple, la classe `can_have_k_name` regroupe la classe `v_museum` (musée) qui est déjà une classe du thème `k_tourism` (tourisme).

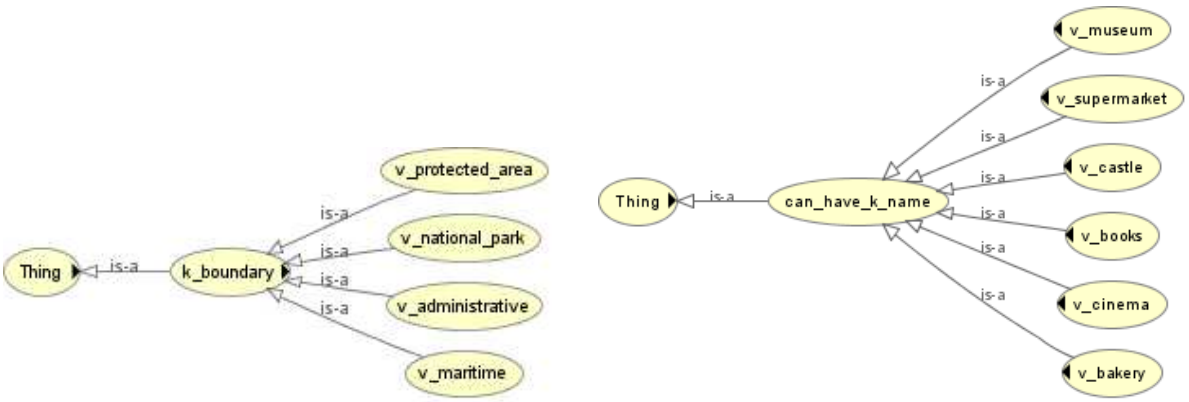



**Figure 74 : (a g.) un exemple des classes OSMonto que nous avons conservé pour notre version française d'OSMonto et, (à d.) un exemple des classes OSMonto que nous avons exclues**

### 2.1.2 La comparaison de deux vocabulaires et la notion de taux de correspondance

Afin de comparer deux vocabulaires, nous éffectuons une comparison par chaines de caracteres (Euzenat et Shvaiko 2007; Atemezing et Troncy 2012) entre les éléments du vocabulaire. Pour cette comparison, nous introduissons la définition de *taux de correspondances* comme le degré de couverture par thème d'une source de vocabulaire par rapport à une autre source de vocabulaire. Nous utilisons cette notion pour quantifier le degré de couverture par thème d'OSMonto-fr par rapport à chacune des sources externes de vocabulaire utilisées par notre processus : Wikipédia, DBpedia, WordNet, IGN. L'objectif est d'établir si les résultats obtenus à chaque étape du processus à partir d'OSMonto-fr, sont satisfaisants ou non. Ces correspondances sont de type 1:1, c'est-à-dire, une classe OSMonto-fr peut correspondre à zéro ou un élément dans l'autre source. Plus précisément, l'Équation 9 exprime la notion de *taux de correspondances* **($TC_{thème}$)** qui est le rapport par thème entre le nombre de classes d'OSMonto-fr trouvant des classes « équivalentes » dans la source externe et le nombre total de classes du thème. Nous utilisons la notation, par exemple de **$TC_{thème\text{-}OSM\text{-}WCG}$** pour désigner le taux de correspondance entre OSMonto-fr et le graphe de catégories de Wikipédia.

$$\textbf{Taux\_de\_correspondance } (TC_{thème}) = \textbf{Nombre\_de\_classes\_thème trouvant\_correspondances} \div \textbf{Nombre\_total\_classes\_thème}$$

**Équation 9: calcule du taux de correspondances entre les classes OSMonto et une source externe de vocabulaire**

Les valeurs obtenues à partir des taux de correspondances d'un thème ($TC_{thème}$) sont interprétées de la façon suivante :

- bien couvert ($TC_{thème} \geq 0,9$),
- assez bien couvert ($0,5 \leq TC_{thème} < 0,9$),
- plus ou moins couvert ($0,2 \leq TC_{thème} < 0,5$),
- mal couvert ($0 < TC_{thème} < 0,2$),
- pas du tout couvert ($TC_{thème} = 0$).

### 2.1.3 Extraction de types de *features* à partir de la base d'extraction Wikipédia

Le *processus de construction d'extraction d'un vocabulaire formel* est lancé à partir des classes des 19 thèmes comprenant les 257 classes OSMonto-fr. Le but est d'extraire des *éléments Wikipédia* (décrits en section III.2.1) de la *base d'extraction Wikipédia*. Dans le cas de non-correspondance entre une classe OSMonto-fr et une catégorie Wikipédia, la méthode prévoit de comparer (également en termes de chaînes de caractères) la classe niveau #1 (thème) correspondante et la catégorie Wikipédia. Par exemple, la classe `abribus` n'a pas une correspondance directe dans le WCG. Néanmoins, le thème auquel elle appartient

`équipement public` possède en effet une correspondance dans le WCG. L'extraction donne 232 classes OSMonto-fr trouvant des correspondances et 25 classes sans correspondance. Le Tableau 10 montre par thème les classes OSMonto-fr n'ayant pas de correspondance, ainsi que les taux de correspondances et la couverture entre OSMonto-fr et le WCG déterminés grâce à l'Équation 9.

| **Thème** | **Classes OSMonto-fr n'ayant pas des correspondances dans le WCG** | **$TC_{thème-OSM-WCG}$** | **Couverture** |
|---|---|---|---|
| Formation végétale | - | 21 ÷ 21 = 1 | Bien couvert |
| Patrimoine | - | 7 ÷ 7 = 1 | |
| Edifices | - | 8 ÷ 8 = 1 | |
| Station de transport | - | 1 ÷ 1 = 1 | |
| Tourisme | - | 15 ÷ 15 = 1 | |
| Transport par câble | - | 4 ÷ 4 = 1 | |
| Frontières | - | 4 ÷ 4 = 1 | |
| Centrale d'électricité | - | 4 ÷ 4 = 1 | |
| Utilisation du sol | - | 1 ÷ 1 = 1 | |
| Equipements publics | `Bistrot`, `Café`, `Clinique` | 70 ÷ 73 = 0,98 | |
| commerces | `Tabac` | 44 ÷ 45 = 0,97 | |
| Routes | `Arrêt de bus` | 12 ÷ 13 = 0,92 | |
| Loisirs | `Terrain de golf` | 15 ÷ 16 = 0,93 | |
| Chemins de fer | `Petite gare de train`, `Arrêt de tramway` | 7 ÷ 9 = 0,77 | Assez bien couvert |
| Cours d'eau | `Rivière`, `Ruisseau` | 7 ÷ 9 = 0,77 | |
| Défense | `Zone de tir`, `Stand de tir` | 4 ÷ 6 = 0,66 | |
| Énergie | `Câbles aériens à haute tension`, `Câbles aériens à basse tension`, `Poteau de support de câbles` | 2 ÷ 5 = 0,40 | Plus ou moins couvert |
| Aviation | `Porte d'embarquement`, `Piste d'atterrissage`, `Voie de circulation`, `Terminal aéroportuaire` | 2 ÷ 6 = 0,33 | |
| Itinéraires | `Ligne de ferry`, `Itinéraire de randonnée en VTT`, `Itinéraire en réseau ferré`, `Piste de ski`, `Ligne de train`, `Ligne de tramway` | 2 ÷ 8 = 0,25 | |

**Tableau 10 : par thème, les classes OSMonto-fr n'ayant pas de correspondance, ainsi que les taux de correspondances et la couverture entre OSMonto-fr et WCG**

Nous remarquons aisément des taux de correspondances élevés pour la plupart des thèmes. Concernant les classes OSMonto-fr ayant des correspondances, nous pouvons faire les remarques suivantes.

Des nombreuses classes OSMonto-fr concernant les thèmes commerce (32 des 45 classes) et chemins de fer (5 des 7 classes) comme par exemple, les classes `laverie` et `passage à niveau` dans OSMonto-fr, ne correspondent pas directement à une catégorie du WCG. En revanche, ces classes, grâce à leur appartenance à des thèmes, possèdent maintenant une correspondance dans le WCG. Par exemple, les classes OSMonto-fr `laverie` et `passage à niveau` correspondent respectivement, aux catégories du WCG : `commerce` et `chemin de fer`.

Très peu de classes OSMonto-fr correspondent à des catégories du WCG qui représentent des « concepts localisés ». Par exemple, les classes OSMonto-fr `synagogue`, `lycée`, et `collège` appartenant au thème équipement public, trouvent respectivement les catégories du WCG suivantes : `synagogue en Tunisie`, `lycée nancéien`, `collège français`. D'autres exemples sont : `mairie` → `mairie d'arrondissement à Paris`, `mémorial` → `mémorial national américain`, `borne frontière` → `borne frontière monument historique (France)`, `réserve naturelle` → `réserve naturelle d'Écosse`, `sommet` → `sommet des Alpes`, `quai` → `quai parisien`, `rivière` → `rivière du Rempart`.

En considérant que notre processus extrait les catégories du WCG de même que ses sous-catégories, elles sont conservées par le processus pour la prochaine étape. Nous observons que peu de classes OSMonto-fr correspondent à des catégories du WCG avec des nombreuses sous-catégories. Par exemple, cela est le cas pour les classes OSMonto-fr suivantes : `grotte`, littoral et zone humide appartenant au thème formation végétale (des extraits sur la Figure 75), de même que pour aqueduc et usine (avec 10 sous-catégories) du thème édifices (des extraits sur la Figure 76).

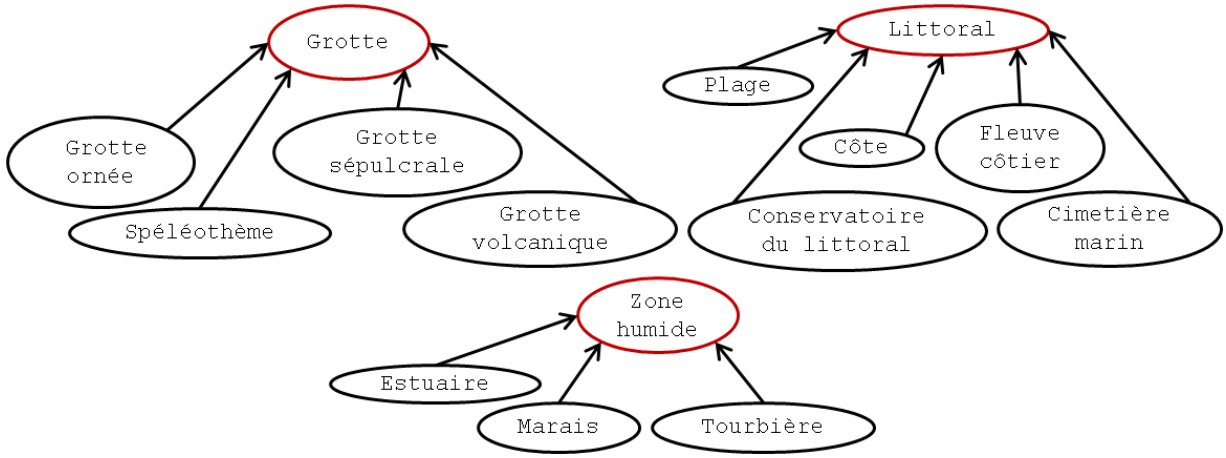



**Figure 75 : des classes du thème « formation végétale » avec ses sous-classes dans le WCG**

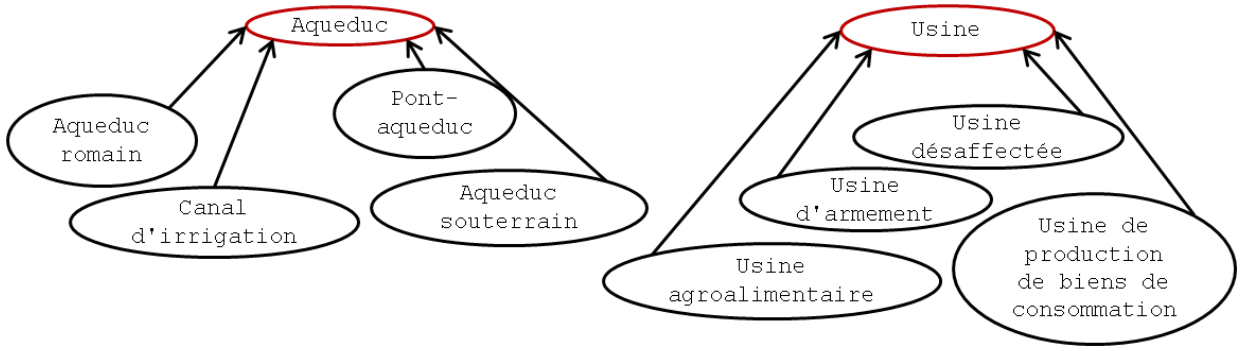



**Figure 76 : des classes du thème « édifices » avec ses sous-classes dans le WCG**

Les 32 classes OSMonto-fr du tableau 1 n'ayant pas de correspondances montrent les limites de notre processus qui effectue des recherches uniquement syntaxiques. Plus précisément, nous observons que la plupart des noms des classes sont plus complexes que ce que le processus ne pouvait s'y attendre, puisque leurs noms sont composés et contiennent des prépositions comme « de » et « en », par exemple, `itinéraire de randonnée en VTT`. Les catégories obtenues à partir de la classe OSMonto-fr comme `tabac`, nous signalent un problème d'homonymie vraisemblable à celui de la résolution d'homonymie concernant les noms de lieux en recherche d'information géographique (Overell et al. 2006). La catégorie du WCG correspondante concerne le produit et non le commerce. Pour la classe OSMonto-fr `Café`, nous observons une nouvelle syntaxe Wikipédia pour clarifier la signification d'une catégorie : `Café_(établissement)`. Il faut noter que dans Wikipédia, l'utilisation de la syntaxe d'homonymie est courante pour les articles mais rare pour les catégories. Pour la classe OSMonto-fr `bistrot`, une autre variation lexicale est utilisée dans Wikipédia : `Brasserie_(restaurant)`. Les autres cas comme `clinique`, `rivière`, et `ruisseau`, sont plus difficiles à expliquer car il apparaît que ce sont des catégories globalement très importantes.

Grâce à cette extraction à partir de la *base d'extraction Wikipédia*, le *processus de construction d'un vocabulaire formel* initialise une instance d'un *Type de Feature* dans *notre catalogue d'éléments de vocabulaire* pour chaque catégorie différente du WCG, c'est-à-dire, 128 types de *features* différents. Par exemple, un type de *feature* `aire protégée` est initialisé et stocké dans le catalogue et puis sur le vocabulaire du contributeur `Cbrando` de la manière suivante :

```
<featureType>
      <nom> Aire protégée </nom>
      <instanciePar> Cbrando </instanciePar>
      <source>
            <nom> graphe de catégories de Wikipédia (WCG)</nom>
            <categorie>
                  <nom> Aire protégée </nom>
                  <parentNodes> Territoire </parentNodes>
                  <parentNodes> Conservation de la nature </parentNodes>
```

```
            </categorie>
        </source>
</featureType>
```

### 2.1.4 Extraction de types de propriétés à partir de la base d'extraction Wikipédia

A partir des 128 types de *features* extraits grâce à OSMonto-fr, le processus procède par extraire des types de propriétés à partir des modèles infobox Wikipédia. Le processus trouve globalement peu de modèles Infobox mais de nombreux types de propriétés (entre 5 et 32) par types de *feature* le sont. Pour illustrer, quelques types de propriétés trouvés concernant le type de feature `centre commercial` sont : `nom`, `propriétaire`, `date d'ouverture`, `fréquentation annuelle`, `chiffre d'affaires annuel`, `slogan`, `site Web`. De la même manière, il existe environ 15 types de propriétés propres au type de *feature* `monument` comme `destination initiale`, `destination actuelle`, `date de construction`, `hauteur`, `ingénieur`, `architecte`, `type`, `style`. Nous observons aussi que la présence d'un modèle infobox implique l'existence d'une catégorie du WCG, mais l'inverse n'est pas vrai.

Le Tableau 11 montre par thème les classes OSMonto-fr (celles qui concernent les 128 types de *features*) ayant des correspondances dans les modèles infobox Wikipédia, ainsi que le nombre total de types de propriétés trouvés par thème, les taux de correspondances et la couverture entre OSMonto-fr et les modèles infobox Wikipédia déterminés grâce à l'Équation 9.

| **Thème** | **Classes OSMonto-fr ayant des correspondances dans les modèles infobox de Wikipédia** | **Nombre total de types de propriété par thème** | **$TC_{thème\text{-}OSM\text{-}Infoboxes}$** | **Couverture** |
|---|---|---|---|---|
| Utilisation du sol | `forêt` | 21 | 1 ÷ 1 = 1 | Bien couvert |
| Station de transport | `station de métro` | 18 | 1 ÷ 1 = 1 | |
| Patrimoine | `site archéologique`, `château`, `monument`, `tombeau historique`. | 87 | 4 ÷ 7 = 0,57 | Assez bien couvert |
| Loisirs | `arène`, `terrain de golf`, `golf miniature`, `parc`, `terrain de sport`, `aire de jeu`, `centre sportif`, `stade`, `terrain de course` | 209 | 9 ÷ 16 = 0,56 | |
| Cours d'eau | `canal`, `barrage`, `fossé`, `égout`, `chute d'eau` | 121 | 5 ÷ 9 = 0,55 | |
| Frontière | `frontière`, `aire protégé` | 45 | 2 ÷ 4 = 0,5 | |
| Tourisme | `refuge de montagne`, `œuvre d'art`, `hôtel`, | 124 | 6 ÷ 15 = 0,4 | Plus ou moins |

| | | | | |
|---|---|---|---|---|
| | musée, parc d'attractions, parc zoologique | | | couvert |
| Routes | aménagement cyclable, échangeur autoroutier, sentier, route nationale, circuit automobile, route départementale | 179 | 7 ÷ 13 = 0,36 | |
| Formation végétale | plage, littoral, montagne, vallée, lac, zone humide | 113 | 6 ÷ 21 = 0,28 | |
| Chemins de fer | chemin de fer, gare ferroviaire | 51 | 2 ÷ 7 = 0,28 | |
| Édifices | aqueduc, usine | 35 | 2 ÷ 8 = 0,25 | |
| Centrale d'électricité | centrale nucléaire | 25 | 1 ÷ 4 = 0,25 | |
| Transport par câble | remontée mécanique | 26 | 1 ÷ 4 = 0,25 | |
| Aviation | Aéroport | 29 | 1 ÷ 6 = 0,16 | Mal couvert |
| Itinéraire | sentier de randonnée | 17 | 1 ÷ 8 = 0,125 | |
| Équipement public | aéroport, banque, hôpital, bibliothèque, église, temple, restaurant, salle de spectacle | 137 | 8 ÷ 73 = 0,10 | |
| Commerce | centre commercial | 16 | 1 ÷ 45 = 0,02 | |

**Tableau 11: par thème, les classes OSMonto-fr ayant des correspondances, ainsi que le nombre total de types de propriétés par thème, les taux de correspondances et la couverture entre OSMonto-fr et les modèles infobox Wikipédia**

Les thèmes non-couverts sont les thèmes énergie et défense.

Grâce à cette extraction, le *processus de construction d'un vocabulaire formel* initialise, dans *notre catalogue d'éléments de vocabulaire*, 1253 instances de type de propriétés correspondantes à nos 128 types de features. Par exemple, un type de propriété `administration` est initialisé et stocké dans le catalogue, puis sur le vocabulaire du contributeur `Cbrando` de la manière suivante :

```
<proprieteType>
      <nom> administration </nom>
      <nomFeatureType> Aire protégée </nomFeatureType>
      <instanciePar> Cbrando </instanciePar>
      <source>
            <nom> Modèles Infobox Wikipédia </nom>
            <modeleInfobox>
```

```
                    <nom> Aire protégée </nom>
                    <champ> administration </champ>
                    …
                </modeleInfobox>
            </source>
</proprieteType>
```

Nous remarquons que certains types de propriétés peuvent être considérés comme des types de relations. Par exemple, le type de propriété `villes principales` concernant le type de *feature* `Route nationale` pourrait être considéré comme un type de relation entre `route nationale` et un type de feature `Ville`. Néanmoins, il n'est pas possible pour un processus de connaître ces informations automatiquement à partir des modèles infobox Wikipédia. Pour cette raison, nous les avons gardés comme des types de propriétés.

### 2.1.5 Extraction de types de propriétés et types de relations à partir de la base d'extraction DBpedia

Le *processus de construction d'extraction d'un vocabulaire formel* extrait des *éléments DBpedia* à partir des 128 types de features trouvés grâce à OSMonto-fr. Pour chaque classe OSMonto-fr correspondant à un type de feature, le processus interroge les classes de l'ontologie DBpedia afin d'extraire des types de propriétés et des types de relations dans lesquelles au moins un des 128 types de features est impliqué.

Le Tableau 12 montre par thème, les classes OSMonto-fr (celles qui concernent les 128 types de *features*) ayant des correspondances dans DBpedia, ainsi que le nombre total de types de propriétés et de types de relations trouvés par thème, les taux de correspondances et la couverture entre OSMonto-fr et DBpedia déterminés grâce à l'Équation 9.

| **Thème** | **Classes OSMonto-fr ayant des correspondances dans DBpedia** | **Nombre total de types de propriétés par thème** | **Nombre total de types de relations par thème** | **$TC_{thème-OSM-DBpedia}$** | **Couverture** |
|---|---|---|---|---|---|
| Station de transport | `station de métro` | 15 | 1 | 1 ÷ 1 = 1 | Bien couvert |
| Frontières | `parc national` et `aire protégé` | 12 | 4 | 2 ÷ 4 = 0,5 | Assez bien couvert |
| Cours d'eau | `canal artificiel`, `ruisseau` et `rivière` | 17 | 24 | 3 ÷ 9 = 0,33 | Plus ou moins couvert |
| Routes | `route national`, `route départementale`, `route communal` et | 56 | 32 | 4 ÷ 13 = 0,30 | |

| | | | | | |
|---|---|---|---|---|---|
| | voie rapide | | | | |
| Patrimoine | borne frontière et monument | 28 | 5 | 2 ÷ 7 = 0,28 | |
| Loisirs | arènes, parc, stade et parc aquatique | 44 | 5 | 4 ÷ 16 = 0,25 | |
| Tourisme | hôtel, musée et parc zoologique | 32 | 7 | 3 ÷ 15 = 0,2 | |
| Aviation | aéroport | 12 | 4 | 1 ÷ 6 = 0,16 | Mal couvert |
| Formation végétale | montagne, lac, et volcan | 8 | 19 | 3 ÷ 21 = 0,14 | |
| équipement public | aéroport, hôpital, bibliothèque, restaurant, école primaire, refuge de montagne, théâtre, université, et établissement d'enseignement supérieur non-universitaire | 90 | 45 | 9 ÷ 73 = 0,12 | |
| Commerce | centre commercial | 4 | 2 | 1 ÷ 45 = 0,02 | |

**Tableau 12 : par thème, les classes OSMonto ayant des correspondances, le nombre total de types de propriétés et les types de relations trouvés ainsi que les taux de correspondances entre les classes OSMonto et les classes de l'ontologie DBpedia**

Les thèmes non-couverts sont les suivants : chemins de fer, itinéraire, édifice, défense, énergie, transport par câble, centrale d'électricité, et utilisation du sol.

Nous observons que l'extraction trouve globalement de nombreux types de propriétés et de relations concernant peu de classes OSMonto-fr. Autrement dit, l'ontologie DBpedia décrit très exhaustivement peu de concepts dans OSMonto-fr. Pour une classe OSMonto-fr, il est possible de trouver entre 1 et 33 types de propriétés, et entre 1 et 21 types de relations. Pour illustrer, le Tableau 13 montre quelques types de propriétés et types de relations DBpedia trouvés à partir des classes OSMonto-fr.

| **Classe OSMonto** | **Types de propriétés et types de relations** |
|---|---|
| aéroport | operateur, plate-forme de correspondance aéroportuaire (*hub*), longitude des pistes d'atterrissage |

| | |
|---|---|
| musée | superficie d'étages, nombre d'étages, conservateur, propriétaire |
| école primaire | cours offerts, classement, personnel, nombre moyen d'élèves par cours, couleurs officielles de l'école, code de l'école |
| volcan | année d'éruption |
| université | recteur, nombre de doctorants, nombre d'employés académique |
| lac | congelé, longueur du rive, embouchure du lac |

**Tableau 13 : types de propriétés et types de relations DBpedia trouvés par le processus à partir des classes OSMonto-fr**

Nous notons également une différence de couverture entre le vocabulaire Wikipédia (voir le Tableau 10) et DBpedia (voir le Tableau 12). Globalement, le WCG possède une meilleure couverture de thèmes OSMonto-fr que l'ontologie DBpedia. Par exemple, des thèmes comme transport par câble et centrales d'électricité sont couverts par Wikipédia et non couverts par DBpedia. Le fait de conserver le WCG en plus de DBpedia était nécessaire car de nombreux concepts sont utilisés par Wikipédia et n'ont pas encore été incorporés dans DBpedia, par exemple, les classes `refuge de montagne` et `centre commercial`. Plus récemment, il est possible d'éditer l'ontologie DBpedia[68], et ce fait peut influencer l'évolution de cette ontologie dans le futur proche.

Grâce à cette extraction à partir de la *base d'extraction DBpedia*, le *processus de construction d'un vocabulaire formel* initialise, dans *notre catalogue d'éléments de vocabulaire*, 346 instances de *types de propriétés* et 147 *types de relations* répartis sur nos 128 types de features déjà existants ainsi que sur 14 nouveaux types de features qui ont été ajoutés par le processus dans le catalogue. Ces nouveaux types de features correspondant à des classes DBpedia sont les suivants : `Personne`, `Organisation`, `Pays`, `Compagnie aérienne`, `Entreprise`, `Lieu habité`, `Lieu`, `Cité`, `Chaîne de montagne`, `Etendu d'eau`, `Architecte`, `Infrastructure`, `Artiste` et `Ile`. Par exemple, un type de relation `ville proche` entre les types de features `Aire protégée` et `Lieu habité` est initialisé et stocké dans le catalogue et sur le vocabulaire du contributeur `Cbrando` de la manière suivante :

```
<relationType>
      <nom> ville proche </nom>
      <nomFeatureType1> Aire protégée </nomFeatureType1>
      <nomFeatureType2> Lieu habité </nomFeatureType2>
      <instanciePar> Cbrando </instanciePar>
      <source>
            <nom> Ontologie DBpedia </nom>
            <typeRelationOnto>
                  <nom> ville proche </nom>
                  <domaine> lieu </domaine>
```

[68] http://mappings.dbpedia.org/index.php/Ontology_Editing

```
                <range> lieu habité </range>
            </typeRelationOnto>
        </source>
</relationType>
```

Nous remarquons que certains types de relations trouvés possèdent un contexte non-spatial comme par exemple, le type de relation `opérateur` entre `Station de métro` et `Organisation`. Nous les avons gardés dans le catalogue car ces types de relations peuvent être pertinents pour des contenus thématiques. Nous trouvons aussi une certaine quantité de types de propriétés redondantes par rapport à des types de propriétés obtenus auparavant dans les modèles infobox Wikipédia. En effet, l'ontologie DBpedia a été construite à partir des infoboxes de Wikipédia, ce comportement est donc attendu. Nous avons décidé de garder les réponses obtenues pour cette initialisation du catalogue afin de ne pas risquer d'enlever des faux positifs.

### 2.1.6 Extraction de types de relations à partir de la base d'extraction WordNet

Le *processus de construction d'extraction d'un vocabulaire formel* extrait des *éléments WordNet*, à partir des 128 types de *features* trouvés grâce à OSMonto-fr. Pour chaque classe OSMonto-fr correspondant à un type de *feature* de notre catalogue, le processus extrait des relations lexicales WordNet dont au moins un de ses littéraux correspondent à un type de feature.

Le Tableau 14 montre par thème, les classes OSMonto-fr (celles qui concernent les 128 types de *features*) ayant des correspondances dans WordNet, ainsi que le nombre total de types de relations trouvés par thème, les taux de correspondances et la couverture entre OSMonto-fr et WordNet déterminés grâce à l'Équation 9.

| **Thème** | **Classes OSMonto-fr ayant des correspondances dans WordNet** | **Nombre total de types de relations par thème** | $TC_{thème\text{-}OSM\text{-}WordNet}$ | **Couverture** |
|---|---|---|---|---|
| Utilisation du sol | `Forêt` | 2 | 1 ÷ 1 = 1 | Bien couvert |
| Chemins de fer | `Gare ferroviaire, Arrêt de tramway, Passage à niveau, Portion de voie désaffectée, Passage piéton, Ancien voie de chemin de fer (voie ferrée)` | 6 | 6 ÷ 9 = 0,66 | Assez bien couvert |

| | | | | |
|---|---|---|---|---|
| Formation végétale | Baie (mer et lac), Sommet (chemin), Arbre remarquable (bois), Bois (sous-bois) | 8 | 8 ÷ 21 = 0,38 | Plus ou moins couvert |
| Commerce | Fleuriste (place de marché) | 1 | 1 ÷ 45 = 0,02 | Mal couvert |

**Tableau 14 : par thème, les classes OSMonto-fr ayant des correspondances dans WordNet, le nombre total de types de relations trouvés par thème ainsi que les taux de correspondances et la couverture entre OSMonto-fr et WordNet**

Nous remarquons un mauvais degré de couverture entre WordNet et OSMonto-fr car le reste des 16 thèmes OSMonto-fr ne sont pas du tout couverts par WordNet. De plus, le processus trouve globalement peu de types de relations WordNet pour peu de thèmes OSMonto-fr. Autrement dit, la base de données lexicale WordNet ne décrit, ni exhaustivement ni en ampleur ces classes en termes de types de relation. Le Tableau 15 montre quelques types de propriété et types de relations WordNet trouvés par le processus, à partir des classes OSMonto-fr. Nous observons des types de relations inter-thèmes, par exemple, le type de relation entre `Fleuriste` du thème commerce et `place de marché` du thème équipement public. Nous remarquons également des types de relations intra-thèmes, par exemple, le type de relation entre `Baie` et `Lac` du thème formation végétale.

| Classe OSMonto-fr #1 (et son thème) impliquée | Types de relation WordNet | Classe OSMonto-fr #2 (et son thème) impliquée |
|---|---|---|
| Fleuriste (commerce) | « compose un marché » | Place de marché (équipement public) |
| Baie (formation végétale) | « compose un lac » | Lac (formation végétale) |
| Forêt (utilisation du sol) | « compose d'arbres » | Arbre remarquable (utilisation du sol) |
| Forêt (utilisation du sol) | « compose de sous-bois » | Bois (formation végétale) |

**Tableau 15 : types de relation WordNet, inter-thèmes et intra-thèmes, trouvés par le processus à partir des classes OSMonto-fr**

Grâce à cette extraction à partir de la *base d'extraction WordNet*, le *processus de construction d'un vocabulaire formel* initialise dans *notre catalogue d'éléments de vocabulaire*, uniquement 11 instances de *types de relations* répartis sur nos 128 types de features déjà existants. Par exemple, un type de relation `compose arbres` entre les types de features `Forêt` et `Arbre remarquable` est initialisé et stocké dans le catalogue, puis sur le vocabulaire du contributeur `Cbrando` de la manière suivante :

```
<relationType>
      <nom> compose arbres </nom>
```

```
      <nomFeatureType1> Forêt </nomFeatureType1>
      <nomFeatureType2> Arbre remarquable </nomFeatureType2>
      <instanciePar> Cbrando </instanciePar>
      <source>
            <nom> Base de données lexicale WordNet </nom>
            <typeRelationWordNet>
                  <type> Méronymie </type>
                  <literal1> forêt </literal1>
                  <literal2> arbre </literal2>
            </typeRelationOntoWordNet>
      </source>
</relationType>
```

D'autres types de relations possibles ont été trouvés comme par exemple la relation « une écluse peut composer un canal », la relation « un carrefour compose une route ». Dans ces cas, des classes `écluse` et `route` sont bien des classes OSMonto-fr. Cependant, des classes comme `carrefour` n'ont pas été définies dans OSMonto-fr (elles ne le sont pas non plus dans la version anglaise originale OSMonto). Ces types de relations ne sont donc pas initialisés dans le catalogue d'éléments de vocabulaire car le processus n'a pas assez d'information de WordNet pour assurer que ces littéraux pourraient correspondre à des types de *features*. Néanmoins, comme nous l'avons expliqué dans la section III.2.3, le processus garde également ces types de relations pour la prochaine étape, afin de rechercher des liens avec les spécifications IGN.

### 2.1.7 Extraction des classes de la base d'extraction IGN afin de créer des types de relations avec des types de feature du vocabulaire

Le *processus de construction d'extraction d'un vocabulaire formel* extrait des *éléments IGN*, à partir des 128 types de *features* trouvés grâce à OSMonto-fr de même qu'à partir des 147 types de relations et des 11 types de relations obtenus par le processus respectivement via DBpedia et WordNet.

Le Tableau 16 montre par thème les classes OSMonto-fr (celles qui concernent les 128 types de *features*) ayant des correspondances avec les spécifications IGN, ainsi que le nombre total de types de relations par thème, les taux de correspondances et la couverture entre OSMonto-fr et WordNet déterminés grâce à l'Équation 9.

| **Thème** | **Classes OSMonto-fr ayant des correspondances dans les spécifications IGN** | **Nombre total de types de relations par thème** | $TC_{thème\text{-}OSM\text{-}IGN}$ | **Couverture** |
|---|---|---|---|---|
| Utilisation de sol | `Forêt` | 3 | 1 ÷ 1 = 1 | Bien couvert |

| | | | | |
|---|---|---|---|---|
| Cours d'eau | `Fleuve, canal artificiel, rivière` | 3 | 3 ÷ 7 = 0,42 | Plus ou moins couvert |
| Routes | `Autoroute, route nationale, route départementale, route communal, voie rapide` | 9 | 5 ÷ 13 = 0,38 | |
| Chemins de fer | `Passage piéton, passage à niveau, arrêt de tramway` | 3 | 3 ÷ 9 = 0,33 | |
| Frontières | `Aire protégée` | 1 | 1÷ 4 = 0,25 | |
| Formation végétale | `Lac, montagne` | 10 | 2 ÷ 21 = 0,09 | Mal couvert |

**Tableau 16 : par thème, les classes OSMonto-fr ayant des correspondances avec les spécifications IGN via des types de relations de notre catalogue, le nombre de types de relations, les taux de correspondances et la couverture**

Les thèmes non-couverts sont les suivants : transport par câble, station de transport, édifices, loisirs, patrimoine, énergie, tourisme, centrale d'électricité, défense, aviation, équipement public, itinéraires, équipements publics.

Le processus ne trouve globalement pas une bonne quantité de types de relations entre le vocabulaire construit et les spécifications de référence. Le Tableau 17 montre des types de relations trouvés par le processus, les classes OSMonto-fr et les classes IGN impliquées. Le fait de trouver peu de types de relations avec les spécifications IGN est une conséquence directe de ne pas avoir trouvé assez de types de relations WordNet ou DBpedia. Certes, le processus disposait de 147 types de relations DBpedia. Néanmoins, ces types de relations ne sont pas tous de nature « spatiale » : il s'agit de relations concernant des personnes ou des organisations. Une autre raison qui peut expliquer ces résultats est la nature des types de relations trouvés dans WordNet car elles concernent des concepts avec un très faible niveau de détail (ex : l'aéroport est décomposé en hangar, tour de contrôle, etc.) par rapport au niveau de detail des bases de données IGN. De plus, les ressources formelles de spécifications, c'est-à-dire l'ontologie et la description du schéma produit dont le processus se sert ne sont pas assez riches en relations. En effet, le processus trouve de nombreux « liens » entre une classe OSMonto-fr et une classe IGN mais le type de ce lien est inconnu. Par exemple, il existe un lien entre `Parc national` et `Tronçon de route` mais le processus ne peut pas le nommer. Certains d'eux sont clairement des liens de spécialisation, par exemple, `Canalisation` et `Aqueduc`, mais le processus n'est pas capable de faire cette inférence. D'autres exemples sont : `Usine` et `Bâtiment`, `Traitement de l'eau` et `Bassin d'épuration`.

| **Classes OSMonto-fr (et sont thème) impliquée** | **Type de relation (WordNet ou DBpedia)** | **Classe IGN impliquée** |
|---|---|---|
| `Forêt (utilisation du sol)` | « peut composer » (WordNet) | `Zone arborée` |
| `Forêt (utilisation du sol)` | « peut composer » (WordNet) | `Lieu-dit non habité` |

| | | |
|---|---|---|
| Forêt (utilisation du sol) | « peut être sous-classe » (WordNet) | massif boisé |
| Fleuve (cours d'eau) | « peut être composé » (WordNet) | Tronçon cours d'eau |
| Rivière (cours d'eau) | « source rencontre » (DBpedia) | Montagne |
| Rivière (cours d'eau) | «bras de rivière » (DBpedia) | Tronçon de cours d'eau |
| Autoroute (routes) | « peut composer » (WordNet) | Voie mère de branchement |
| Autoroute (routes) | « peut composer » (WordNet) | Surface de route |
| Autoroute (routes) | « croise » (DBpedia) | Tronçon de route |
| Lac (formation végétale) | « sortie d'eau » (DBpedia) | Tronçon de cours d'eau |
| Lac (formation végétale) | « entrée d'eau » (DBpedia) | Tronçon de cours d'eau |
| Passage à niveau (chemins de fer) | « peut composer » (WordNet) | Tronçon de voie ferrée |
| Aire protégée (frontières) | « ville la plus proche » (DBpedia) | Lieu-dit habité |

**Tableau 17 : des types de relations sur lesquels il est possible de raccrocher le vocabulaire construit et les spécifications IGN**

Grâce à cette extraction à partir de la *base d'extraction IGN*, le *processus de construction d'un vocabulaire formel* initialise, dans *notre catalogue d'éléments de vocabulaire*, 29 instances de *types de relations* répartis sur nos 128 types de features déjà existants et des types de feature IGN. Par exemple, un type de relation `entrée d'eau` entre les types de features `Lac` et `Tronçon de cours d'eau` de l'IGN est initialisé et stocké dans le catalogue ainsi que sur le vocabulaire du contributeur `Cbrando` de la manière suivante :

```
<relationType>
      <nom> Entrée d'eau </nom>
      <nomFeatureType1> Lac </nomFeatureType1>
      <nomFeatureType2> Tronçon de cours d'eau </nomFeatureType2>
      <estFeatureTypeReference> Tronçon de cours d'eau
      </estFeatureTypeReference>
      <instanciePar> Cbrando </instanciePar>
      <source>
            <nom> Spécifications IGN </nom>
            <classe>
                  <nom>Tronçon de cours d'eau</nom>
                  <description> Portion de cours d'eau, réel ou fictif,
                  permanent ou temporaire, naturel ou artificiel, homogène
                  pour l'ensemble des attributs qui la concernent, et qui
                  n'inclut pas de confluent.
```

```
                </description>
            </classe>
        </source>
</relationType>
```

### 2.1.8 Construction d'un catalogue d'éléments de vocabulaire

Comme nous l'avons signalé à chaque étape, le processus a semi-automatiquement initialisé un catalogue d'éléments de vocabulaire composé de 142 types de features (128 obtenus à partir du WCG et 14 recréés à partir de DBpedia), 1599 types de propriétés (1253 obtenus à partir des modèles infobox Wikipédia et 346 à partir de DBpedia), et 187 types de relations (147 sont obtenus à partir de DBpedia, 11 de WordNet et 29 sont des types de relations avec des spécifications de référence de l'IGN) (voir le Tableau 18).

| **# Types de *features*** | **# Types de propriétés** | **# Types de relations** | |
|---|---|---|---|
| | | **intra-contenu** | **Vers les spécifications IGN** |
| 128 | 1599 | 158 | 29 |

**Tableau 18 : les éléments de vocabulaire initialisés dans notre catalogue grâce au processus de construction d'un vocabulaire formel et OSMonto-fr**

Nous pouvons également décrire ce catalogue en regard des vocabulaires externes utilisés. Nous considérons les thèmes comme ayant une « bonne couverture globale », ceux qui appartiennent aux catégories suivantes : « bien couvert », « assez bien couvert » et « plus ou moins couvert ». Au contraire, les thèmes ayant une « mauvaise couverture globale » sont ceux qui appartiennent aux catégories suivantes : « mal couvert » et « pas du tout couvert ». Le Tableau 19 illustre les couvertures globales obtenues par thème.

| **Thème** | **WCG** | **Infoboxes** | **Dbpedia** | **Wordnet** | **IGN** |
|---|---|---|---|---|---|
| Equipements publics | + | + | — | — | — |
| Commerces | + | + | — | — | — |
| Formation végétale | + | + | — | — | — |
| Loisirs | + | + | + | — | — |
| Tourisme | + | + | + | — | — |
| Routes | + | + | + | — | + |
| Cours d'eau | + | + | + | — | + |
| Chemins de fer | + | + | — | + | + |
| Edifices | + | + | — | — | — |
| Itinéraires | + | + | — | — | — |
| Patrimoine | + | + | + | — | — |
| Energie | + | — | — | — | — |
| Défense | + | — | — | — | — |

| | | | | | |
|---|---|---|---|---|---|
| Aviation | + | + | — | — | — |
| Frontières | + | + | + | — | + |
| Transports par câble | + | + | — | — | — |
| Centrales d'électricité | + | + | — | — | — |
| Utilisation du sol | + | + | — | + | + |
| Stations de transport | + | + | + | — | — |

**Tableau 19 : couverture globale des thèmes du vocabulaire construit par notre méthode, + représente une bonne couverture et — une mauvaise couverture**

Ce catalogue possède globalement une bonne couverture globale pour un nombre appréciable de thèmes de notre vocabulaire en termes de types de *features* et de types de propriété. La couverture au niveau des types de relations (et vers les spécifications IGN) est moins bonne. Néanmoins, nous considérons que ces types de relations sont intéressants pour peupler un catalogue afin qu'il soit enrichi au fur et à mesure par des utilisateurs. En particulier, les thèmes cours d'eau, routes, chemins de fer, frontières et utilisation du sol sont globalement bien couverts, notamment par les spécifications IGN. La couverture des thèmes loisirs, tourisme, patrimoine, stations de transport est acceptable. Malheureusement, les thèmes suivants possèdent une couverture globale insuffisante : équipement publics, commerces, formation végétale, édifices, itinéraires, énergie, défense, aviation, transport par câble et centrales d'électricité.

Les nombreuses catégories Wikipédia possèdent un bon potentiel pour peupler notre catalogue de types de features. Cependant, il existe certaines fois un article Wikipédia (et non une catégorie) fussant référence à un concept comme `rue piétonne` au lieu d'être considéré comme une catégorie dans le WCG. Tout de même, notons l'amélioration considérable du WCG au niveau de la quantité et de la diversité de catégories par rapport à l'an passé. Certaines catégories comme `Autoroute` n'existaient pas mais sous la forme d'un article pendant la réalisation d'une première version de notre processus. Depuis, cette catégorie `Autoroute` (et beaucoup d'autres) existent à présent dans le WCG avec des sous-catégories comme `Echangeur Autoroutier` et `Pont Autoroutier`.

Les modèles Infobox Wikipédia fournissent une quantité appréciable de types de propriétés potentiels pour les types de feature de notre catalogue. En effet, il faut remarquer la richesse des types de propriété pour les classes appartenant à des thèmes comme patrimoine, cours d'eau, formation végétale, loisirs, tourisme et routes. Par exemple, les classes `montagne` et `stade` possèdent respectivement 23 et 28 types de propriété. Néanmoins, les modèles Infobox ne couvrent pas bien les classes concernant les thèmes équipements publics et commerces. Ce fait est contraire à nos hypothèses initiales : ces thèmes sont « populaires » et donc probablement faciles à décrire pour un contributeur de Wikipédia. Des thèmes plus spécialisés comme l'énergie, la défense ou l'aviation sont également mal couverts, un fait tout à fait compréhensible dû à la complexité de ces thèmes. De plus, il faut remarquer que chaque modèle Infobox correspond globalement à une catégorie dans Wikipédia. Néanmoins, cela

n'était pas le cas pendant le développement de la première version de notre processus. Par exemple, la catégorie `Ville` n'existait pas, mais une Infobox et un article avec ce même nom existaient. A l'heure actuelle, la catégorie `Ville` a été créée. L'amélioration de la correspondance entre les modèles Infobox et les catégories est effectuée au fur et à mesure par les contributeurs de Wikipédia.

Nous remarquons également la couverture insuffisante de DBpedia au niveau des types de propriétés et des types de relations disponibles par type de *feature*. Tout de même, il faut apprécier le fait qu'une quantité considérable de concepts importants comme `Montagne`, `Voie de transport`, `Bâtiment` et `Lac` sont très riches en types de propriétés et en types de relations. Un inconvénient dans DBpedia est l'inexistence de « concepts localisés » comme dans le WCG. Par exemple, la classe `collège français` existe dans le WCG mais pas dans DBpedia. Les types de propriétés et les types de relations dans DBpedia représentent donc des caractéristiques plutôt généralistes sans se focaliser sur une région ou ville du monde.

Nous observons une mauvaise couverture globale au niveau des types de relations dans WordNet. De plus, certaines classes possèdent des types de relation provenant de WordNet qui décrivent les instances correspondantes avec un niveau assez réduit de granularité. Par exemple, la classe `bâtiment` inclus deux types de relations `est composé` avec des instances de `mur` et `cour`. Ce type d'information peut être intéressant en regard de l'utilisation qui en sera effectuée.
Nous remarquons une couverture globale insuffisante au niveau des références vers des spécifications IGN. Néanmoins, le processus est capable d'identifier des « références » sans pouvoir automatiquement les nommer. Ce comportement est considéré comme souhaitable car nous voudrions idéalement donner ces suggestions aux contributeurs afin qu'ils puissent nommer les références inconnues pour améliorer le catalogue et afin également de leur donner plus de liberté.

### 2.1.9 Extraction des types de relations spatiales à partir des articles Wikipédia

Un élément clé de notre approche est de suggérer au contributeur des instances de types de relations pour l'aider à construire son vocabulaire. Néanmoins, comme nous l'avons observé dans la section précédente, le processus n'initialise pas dans le catalogue un nombre suffisant de types de relations. Pour cette raison, nous utilisons la méthode d'*extraction de types de relations spatiales à partir des articles Wikipédia* (voir la section III.2.5). Pour la suite, nous avons choisi des types de *features* liés au thème de la ville : des routes, des églises et des fontaines. Ces éléments ont été initialisés dans notre *catalogue d'éléments de vocabulaire* et construit auparavant à partir d'OSMonto-fr. Il faut signaler que cette étape du processus n'est pas effectuée à la volée car il faut effectuer une phase initiale pour l'identification des instances géographiques. De plus, elle devrait être utilisée pour chaque type de *feature* initialisés par le processus dans le *catalogue*.

Grâce à des pages `Listes` de Wikipédia, Nous avons constitué trois groupes d'instances : 121 fontaines, 137 églises et 5536 routes de Paris. A partir de ces 5794 entités, il est possible d'obtenir de notre *base d'extraction Wikipédia*, les textes des articles dont le titre correspond à un nom de ces entités, par exemple, le texte de l'article `Fontaine Millénaire`. Afin d'extraire des types de relations spatiales, nous avons sélectionné à partir d'un ensemble test d'articles 24 termes correspondant à des types de relations topologiques et d'orientation « probables » pour les types de *features* choisis. Par exemple, `croise` et `longe` sont des types de relations qui concernent souvent des routes. Sur les 24 types de relations prédéfinis, notre méthode extrait, à partir des textes des articles, les relations spatiales entre les 5794 entités. Chaque relation extraite est exprimée comme un triplet `<entité1, étiquette type de relation spatiale, entité2>`.

Par exemple, deux triplets obtenus par la méthode sont listés ci-dessous :

```
<Fontaine Saint-Michel, se situe sur la, Place Saint Michel>
<Fontaine du bassin Soufflot, en face de la, rue Soufflot>
```

Notre méthode propose également une représentation visuelle du graphe correspondant, comme celui la Figure 77.

Ensuite, la méthode calcule la fréquence d'utilisation des relations à partir des triplets. La Figure 78 montre l'histogramme des fréquences construit par notre méthode. Nous observons que les relations d'orientation les plus utilisées sont : `au nord`, `au sud`, la relation métrique la plus utilisée est `à proximité`, et les relations topologiques les plus utilisées sont `croise` et `longe`.

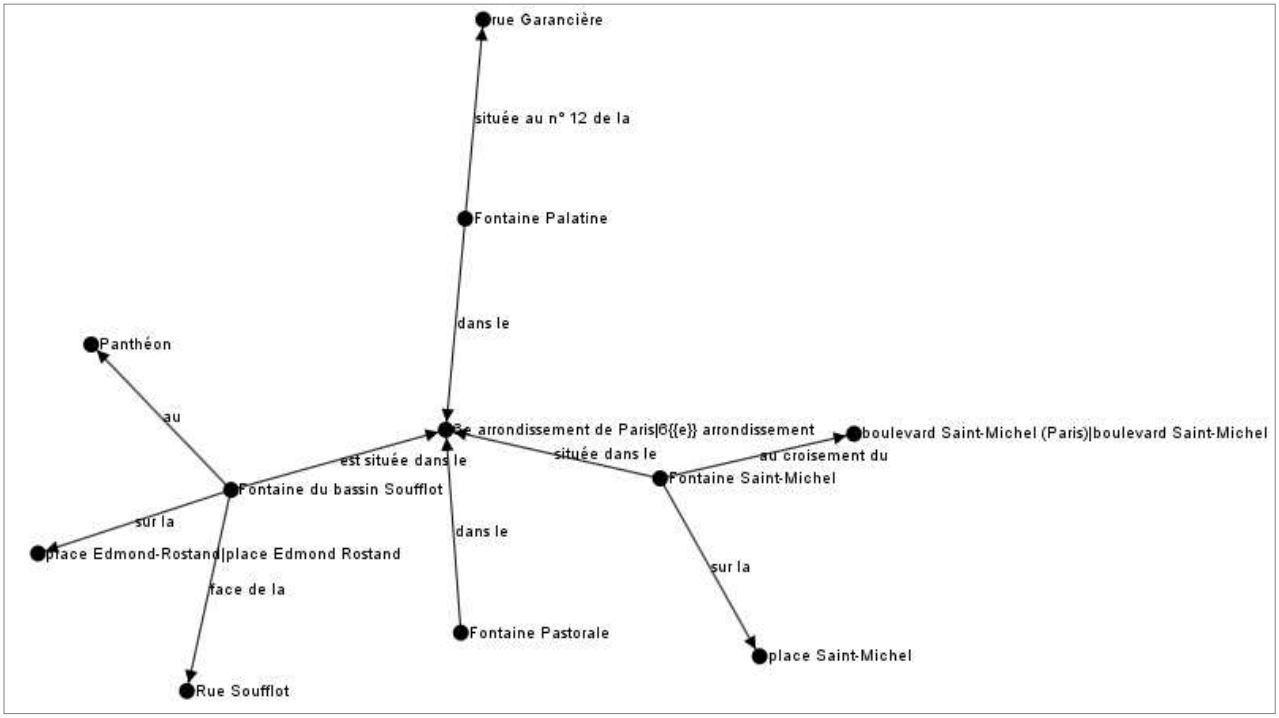



**Figure 77 : un extrait du graphe de relations spatiales construit à partir des fontaines, églises et routes parisiennes**

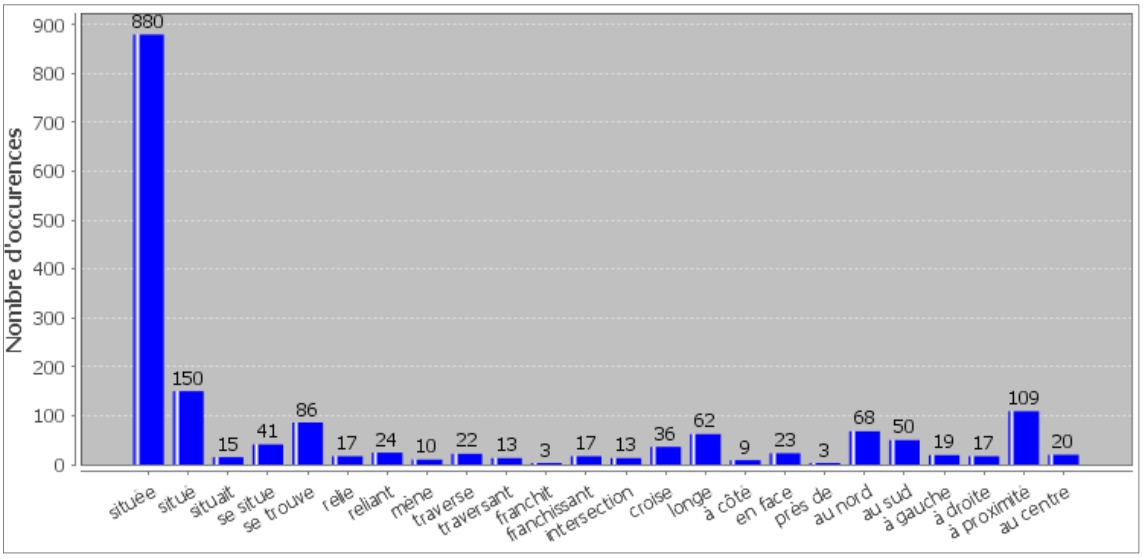



**Figure 78 : fréquence d'utilisation, dans les textes des articles sélectionnés, de relations spatiales entre les 5794 entités, à partir de 24 types de relations prédéfinis**

De plus, la méthode extrait également pour chaque type de relation choisi, par exemple, `croise` et `longe`, les fréquences d'utilisation d'une relation en considérant aussi les types de *features* choisis. Notre méthode produit l'histogramme correspondant sur 4 relations, présenté en Figure 79.

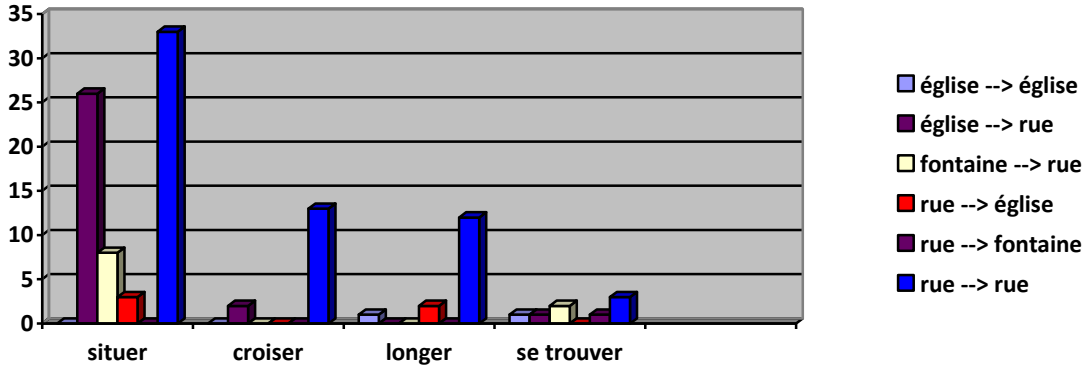



**Figure 79 : fréquence d'utilisation, dans les textes des articles sélectionnés, de relations spatiales entre les 5794 entités, à partir de 24 types de relations prédéfinis et en considérant les types de features choisis**

De cette manière, il a été possible d'initialiser 10 nouveaux types de relations dans notre *catalogue d'éléments de vocabulaire* en considérant uniquement trois types de *features*. Par exemple, un nouveau type de relation `longe` entre des types de *feature* `Routes` est instancié par `Cbrando` et stocké dans le catalogue de la manière suivante :

```
<relationType>
      <nom> longe </nom>
      <nomFeatureType1> Route </nomFeatureType1>
      <nomFeatureType2> Route </nomFeatureType2>
      <instanciePar> Cbrando </instanciePar>
      <source>
            <nom> Articles Wikipédia </nom>
            <catWCG> Fontaines de Paris du 6ème arrondissement </catWCG>
            <catWCG> Fontaines de Paris du 7ème arrondissement </catWCG>
            …
            <catWCG> Voie du 6ème arrondissement de Paris </catWCG>
            …
            <frequenceRelation> 12 </frequenceRelation>
      </source>
</relationType>
```

## 2.2 Construction d'un vocabulaire par des chercheurs en géomatique

Une fois le *catalogue d'éléments de vocabulaire* initialisé, nous avons expérimenté notre *processus d'aide à la construction d'un vocabulaire* ciblé avec des chercheurs du Laboratoire COGIT de l'IGN. Le processus accède au catalogue construit auparavant afin d'encourager la réutilisation des éléments déjà instanciés, et si nécessaire, le processus va rechercher dans les *bases d'extraction*.

### 2.2.1 Description de l'expérience

Le laboratoire COGIT[69] de l'IGN travaille sur les problématiques liées à l'utilisation des données topographiques vectorielles comme l'intégration de données, l'automatisation des processus de généralisation cartographique, la prise en compte et l'évaluation de la qualité des données, la conception de légendes personnalisées, la modélisation spatio-temporelle pour faciliter des analyses sur des dynamiques du territoire comme l'évolution des tissus urbaines et les déplacements animaliers aux différents milieux.

Pour notre expérience, nous avons demandé à quatre chercheurs de ce laboratoire de décrire un contenu collaboratif qu'ils souhaiteraient construire pour répondre à un besoin de données dans leurs travaux de recherche. Ces besoins sont exprimés par des mots clés (possiblement composés) décrivant la nature ou la fonction de la donnée souhaitée. En réponse à ces mots-clés, un processus d'extraction propose des types de *features*, des types de propriétés, et des types de relations, notamment vers des spécifications IGN auxquels raccrocher le contenu collaboratif. Ce test permet aussi d'enrichir et diversifier la portée du catalogue d'éléments de vocabulaire avec des nouveaux thèmes comme par exemple, le tourisme ou les déplacements de faune. Le participant est aussi censé expliquer sur ses choix de sélection. Il peut signaler de nouveaux éléments de vocabulaire dont il a besoin afin de les ajouter manuellement dans le catalogue. Comme nous avons signalé dans la section IV.2.1.8, il arrive que des « références » vers les éléments IGN soient identifiées par le processus, mais il n'est pas possible de les nommer automatiquement. Pour ce test, nous avons demandé aux participants de nommer ces références et d'expliquer leur nature. Ensuite, nous montrons des extraits des vocabulaires construits par les participants à l'aide de notre *processus de construction d'un vocabulaire formel*.

L'intitulé du test proposé aux participants est présenté ci-dessous :

*Je voudrais vous demander de réfléchir à un contenu collaboratif que vous voudriez construire (s'il y avait des outils simples pour cela…) par exemple pour répondre à un besoin de données dans vos travaux de recherche. Il n'y a pas de limites pour le nombre de mots-clés, il est souhaitable de choisir entre 6 à 15 mots-clés. Un mot-clé peut être simple (ex : autoroute) ou composé (ex : ligne de métro). Le test peut durer jusqu'à 25 minutes au total. Il consiste à :*

[69] http://recherche.ign.fr/labos/cogit/accueilCOGIT.php

1) *Préparer des mots-clés (composés ou non) décrivant la nature ou fonction des données souhaitées.*
2) *Lancer le processus à partir des mots-clés. Il sera éventuellement souhaitable de changer certains mots-clés si le processus obtient des réponses vides. En fonction du nombre de mots-clés, le processus peut prendre un temps considérable pour sortir les résultats (pas plus de deux minutes grâce à des indexes).*
3) *Sélectionner les éléments du vocabulaire considérés comme pertinents selon votre besoin.*

*Le but de cette expérience est de tester la pertinence des réponses afin de construire le modèle souhaité. Nous sommes conscients que l'interface développée n'est pas ergonomique car les résultats sortent en XML. Pour cette raison, je vous aiderai à trier les résultats obtenus en XML mais c'est vous qui les choisirez. Ce test, s'il ne porte pas sur l'interface, donne quand même l'opportunité d'avoir des remarques sur l'interface.*

### 2.2.2 Description des résultats

Les thèmes choisis par les participants ont été : les déplacements de faune, les littoraux, les chantiers en ville, et le tourisme rural. Pour chaque thème, nous comptons les éléments du vocabulaire trouvés par le processus dans le catalogue ou dans les bases d'extraction, de même que ceux qui ont été choisis par chaque participant. Certaines fois, même si la réponse obtenue par le processus à partir d'un mot-clé dans le catalogue était fournie, nous avons relancé une recherche forcée à partir de ce mot-clé dans les bases d'extraction. Ensuite, nous décrivons les vocabulaires construits ainsi que les références vers les spécifications IGN par des types de relations. De plus, nous décrivons de nouveaux éléments qui ont été définis à la demande du participant, comme des nouveaux types de features ou des types de relations dont le participant avait besoin dans son contenu.

#### Cas #1 : un vocabulaire sur les déplacements de faune

Le participant s'intéresse au sujet de la modélisation de données géographiques pour l'étude écologique des déplacements de la faune. Le Tableau 20 montre des mots-clés employés par le participant, le nombre d'éléments de vocabulaire trouvés par le processus, soit dans le *catalogue d'éléments de vocabulaire* ou soit dans les *bases d'extraction*, ainsi que le nombre d'éléments signalés pertinents pour le thème du participant. De nouveaux mots-clés ont été suggérés par le participant pendant l'expérience, c'est le cas du thème `végétation`.

| **Mot-clé** | **Nombre total d'éléments de vocabulaire trouvés par le processus** | | **Nombre d'éléments de vocabulaire choisis par le participant** |
|---|---|---|---|
| | **Extraits à partir du catalogue** | **Extraits à partir des bases d'extraction** | |
| Pont | 0 | 88 | 30 |

| | | | |
|---|---|---|---|
| Forêt | 27 | 0 | 16 |
| Autoroute | 8 | 0 | 6 |
| Mammifères | 0 | 24 | 4 |
| Clôture | 0 | 1 | 1 |
| Large | 0 | 4 | 1 |
| Végétation | 0 | 5 | 1 |
| Recouvert végétal | 0 | 0 | 0 |
| Grillage | 0 | 0 | 0 |
| Corridor | 0 | 0 | 0 |
| Passage-faune | 0 | 0 | 0 |
| **Total** | 35 | 122 | 59 |

**Tableau 20 : par mot-clé, le nombre d'éléments pertinents pour le participant concernant les déplacements de faune parmi le total d'éléments obtenus par le processus**

Nous pouvons remarquer la faible présence d'éléments de vocabulaire concernant le thème de l'écologie dans le *catalogue d'éléments de vocabulaire* ainsi que dans les *bases d'extraction*. En effet, l'ontologie OSMonto-fr utilisée pour initialiser ce catalogue, de même que les sources externes de vocabulaire utilisées, ne sont pas ciblées sur le thème écologique. Une évidence est l'absence de réponses à partir des mots-clés : `recouvert végétal`, `grillage`, `corridor` et `passage-faune`. En revanche, les mots-clés `mammifères` et `clôture` fournissent des éléments de vocabulaire signalés pertinents par le participant. Nous observons également la réutilisation des éléments de vocabulaire initialisés grâce à OSMonto-fr, c'est le cas des mots-clés : `forêt` et `autoroute`. Les 59 éléments choisis par le participant sont copiés dans le *catalogue d'éléments de vocabulaire* pour leur réutilisation par les autres participants. Cette étape représente également une phase additionnelle d'enrichissement de notre catalogue avec d'éléments ciblés sur le thème de l'écologie.

La Figure 80 présente un extrait en UML du vocabulaire initialisé par le chercheur intéressé par les déplacements de faune à l'aide de notre *processus de construction d'un vocabulaire formel*. Nous signalons également si l'élément a été trouvé dans le catalogue construit à partir d'OSMonto-fr et sa source.

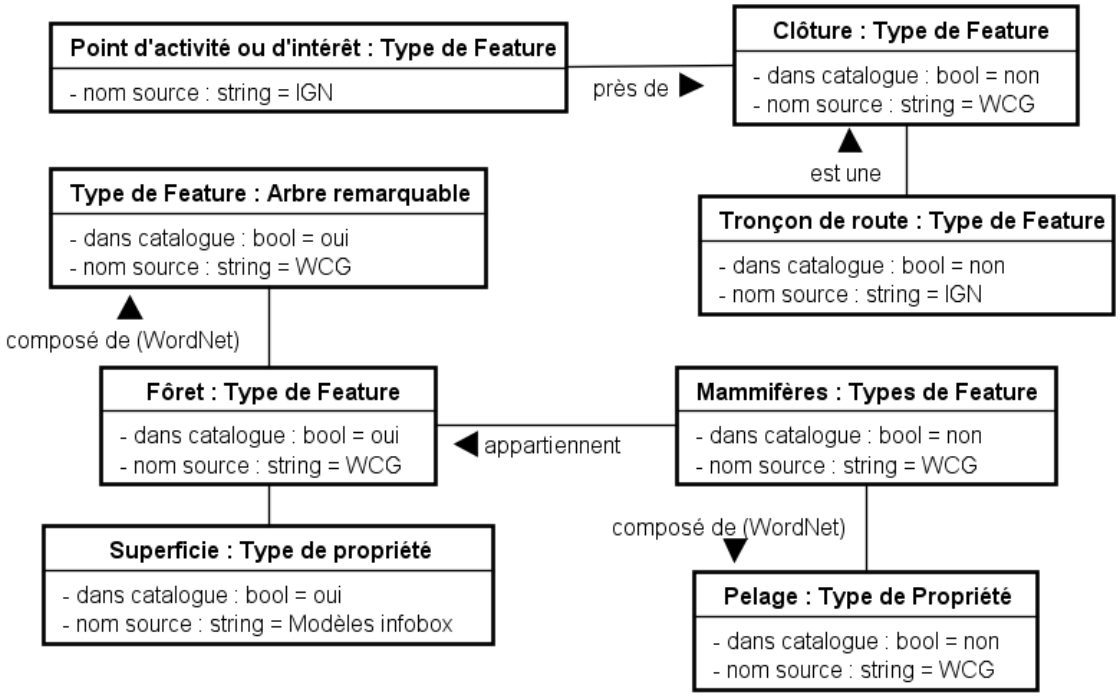



**Figure 80 : extrait de certains éléments de vocabulaire choisis et définis par le chercheur travaillant sur le thème des déplacements de faune**

Le participant s'intéressait à décrire dans son vocabulaire les mammifères et leurs pelages, éléments extraits à la volée grâce à Wikipédia, de même que les forêts et leurs arbres remarquables, éléments existants dans notre catalogue. Ces types de relations ont été établis grâce à WordNet. Un nouveau type de propriété `pelage` a été manuellement créé. Le concept de clôture est très important en écologie, il désigne tout obstacle naturel ou artificiel suivant tout ou partie d'un terrain afin d'empêcher des animaux d'y entrer ou d'en sortir. Le participant voulait créer des références `près de` et `est une` entre respectivement, le type de feature `Clôture` et le type de feature `Point d'activité ou d'intérêt` de l'IGN, et entre `Clôture` et le type de feature `Tronçon de route` de l'IGN. Un nouveau type de relation `appartiennent` entre `Mammifères` et `Clôture`, a aussi été ajouté au vocabulaire à la demande du participant. Par exemple, un nouveau type de relation est défini de cette manière :

```
<relationType>
      <nom> est une </nom>
      <nomFeatureType1> Clôture </nomFeatureType1>
      <nomFeatureType2> Tronçon de route </nomFeatureType2>
      <estFeatureTypeReference> Tronçon de route
      </estFeatureTypeReference>
      <instanciePar> Participant#1 </instanciePar>
      <source>
            <nom> Contributeur Particpant#1 </nom>
      </source>
```

```
</relationType>
```

### Cas #2 : un vocabulaire sur le littoral

Le participant s'intéresse au sujet du littoral. Le Tableau 21 montre les mots-clés employés par le participant, le nombre d'éléments de vocabulaire trouvés par le processus, soit dans le *catalogue d'éléments de vocabulaire*, soit dans les *bases d'extraction*, ainsi que le nombre d'éléments signalés pertinents pour le thème du participant. Un nouveau mot-clé, `plage`, a été suggéré par le participant pendant l'expérience.

| | **Nombre total d'éléments de vocabulaire trouvés par le processus** | | |
|---|---|---|---|
| **Mot-clé** | **Extraits à partir du catalogue** | **Extraits à partir des bases d'extraction** | **Nombre d'éléments de vocabulaire choisis par le participant** |
| Plage | 17 | 0 | 15 |
| Port | 0 | 29 | 13 |
| Mer | 0 | 25 | 7 |
| Phare | 0 | 24 | 6 |
| Digue | 0 | 3 | 3 |
| Rocher | 0 | 7 | 2 |
| Sable | 0 | 5 | 2 |
| Falaise | 1 | 0 | 1 |
| Dune | 1 | 0 | 1 |
| Galet | 0 | 0 | 0 |
| Estran | 0 | 0 | 0 |
| Courbes bathymétriques | 0 | 0 | 0 |
| Mobilier plage | 0 | 0 | 0 |
| **Total** | 19 | 93 | 50 |

**Tableau 21: par mot-clé, le nombre d'éléments pertinents pour le participant concernant les littorales parmi le total d'éléments obtenus par le processus**

Nous pouvons remarquer la faible présence d'éléments de vocabulaire concernant le thème du littoral dans le *catalogue d'éléments de vocabulaire*. En effet, l'ontologie OSMonto-fr utilisée pour initialiser ce catalogue semble ne pas être bien ciblé sur le thème du littoral. En revanche, les bases d'extraction fournissent la plupart des éléments. Dans les deux cas, il n'est pas possible d'extraire des éléments à partir des mots-clés très ciblés sur le thème du littoral, comme : `galet`, `estran` et `courbes bathymétriques`. Nous observons également la faible réutilisation des éléments de vocabulaire initialisés grâce à OSMonto-fr, c'est uniquement le cas du mot-clé : `plage`. Les 50 éléments choisis par le participant sont copiés dans le *catalogue d'éléments de vocabulaire* pour leur réutilisation par les autres participants.

La Figure 81 présente un extrait en UML du vocabulaire initialisé par le chercheur intéressé par le littoral à l'aide de notre *processus de construction d'un vocabulaire formel*. Nous signalons également le cas où l'élément a été trouvé dans le catalogue construit à partir d'OSMonto-fr et sa source.

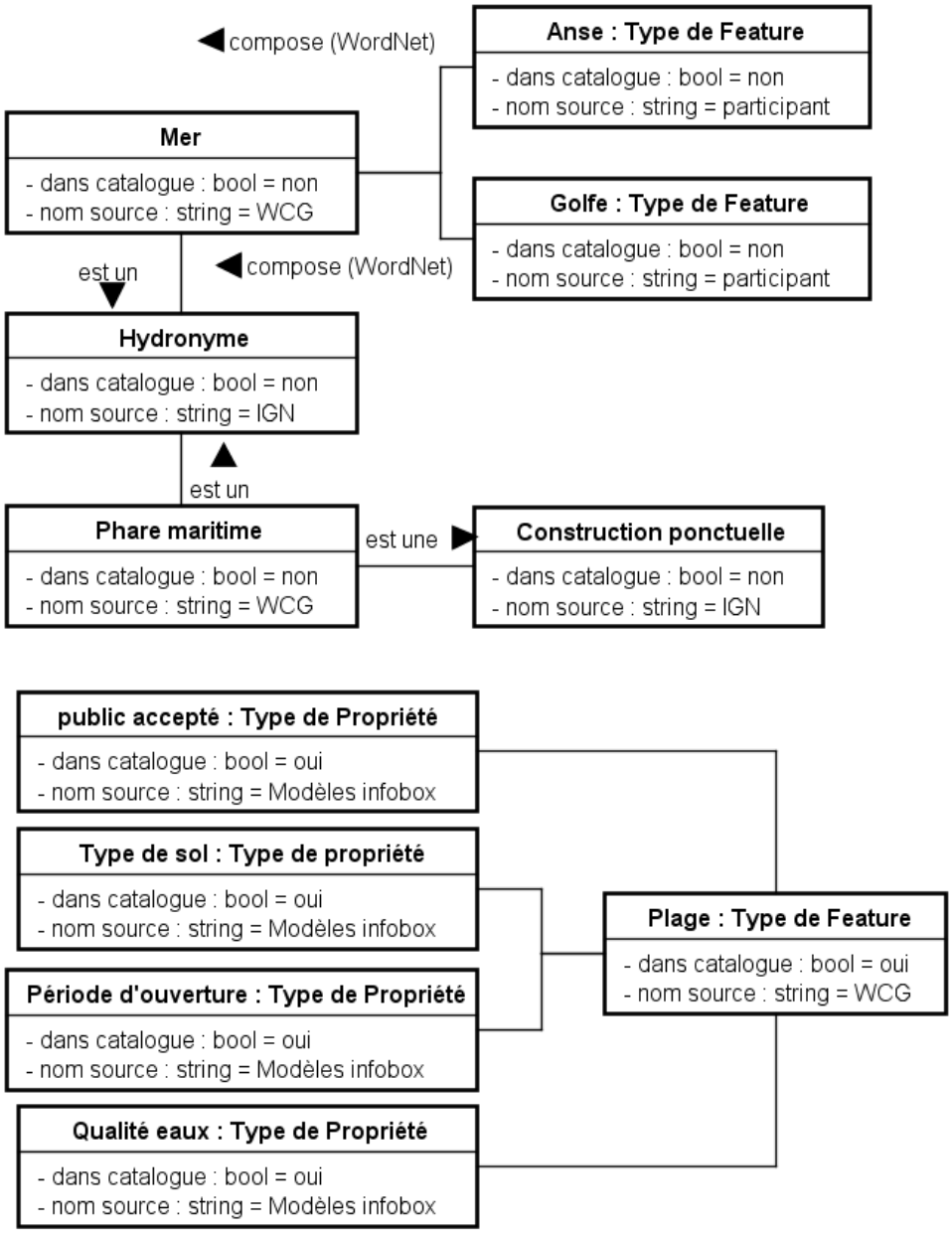



**Figure 81 : extrait de certains éléments de vocabulaire choisis et définis par le chercheur travaillant sur le thème du littoral**

Le participant s'intéressait à décrire dans son vocabulaire, des éléments de paysage qu'il est possible de trouver sur les côtes, comme par exemple les phares. Il voulait également décrire les plages et différentes propriétés qui leur sont associées. Deux nouveaux types de features

ont été créés (`Anse` et `Golf`), qui sont reliés au type de feature `Mer` par des types de relations obtenus grâce à WordNet. Le processus trouve des références vers les spécifications IGN mais elles doivent être nommées par le participant. Par exemple, le type de relation entre `Phare maritime` et `Construction ponctuelle` est signalé important par le participant et est nommé `est une`. C'est également le cas pour les types de relations vers les `Hydronymes`.

**Cas #3 : un vocabulaire sur les chantiers en ville**

Le participant s'intéresse à construire un contenu collaboratif sur les chantiers en ville. Le Tableau 22 montre des mots-clés employés par le participant, le nombre d'éléments de vocabulaire trouvés par le processus, soit dans le *catalogue d'éléments de vocabulaire*, soit dans les *bases d'extraction*, ainsi que le nombre d'éléments signalés pertinents pour le thème du participant.

| | **Nombre total d'éléments de vocabulaire trouvés par le processus** | | |
|---|---|---|---|
| **Mot-clé** | **Extraits à partir du catalogue** | **Extraits à partir des bases d'extraction** | **Nombre d'éléments de vocabulaire choisis par le participant** |
| Bâtiment | 0 | 93 | 14 |
| Rue | 1 | 1 | 2 |
| Gestion de la construction | 0 | 7 | 3 |
| Document d'urbanisme en France | 0 | 1 | 1 |
| Urbanisme | 0 | 2 | 1 |
| Trottoir | 0 | 0 | 0 |
| Travaux | 0 | 0 | 0 |
| Bruit | 0 | 1 | 0 |
| Circulation | 0 | 0 | 0 |
| Parcelle cadastrale | 0 | 0 | 0 |
| Permis de construire | 0 | 0 | 0 |
| **Total** | 1 | 105 | 21 |

**Tableau 22 : par mot-clé, le nombre d'éléments pertinents concernant les chantiers parmi le total d'éléments obtenus par la méthode**

Nous pouvons remarquer l'absence d'éléments de vocabulaire concernant le thème des chantiers dans le *catalogue d'éléments de vocabulaire*. En effet, l'ontologie OSMonto-fr utilisée pour initialiser ce catalogue semble ne pas être ciblée sur le thème des chantiers. En revanche, les bases d'extraction fournissent la presque totalité des éléments choisis par le participant. Dans les deux cas, il n'est pas possible de trouver des éléments à partir de la moitié des mots-clés fournis par le participant, en particulier `travaux`, `circulation`, `parcelle`

`cadastrale` et `permis de construire`. Nous observons également la non-réutilisation des éléments de vocabulaire initialisés grâce à OSMonto-fr, c'est uniquement le cas du mot-clé `rue`. Les 21 éléments choisis par le participant sont copiés dans le *catalogue d'éléments de vocabulaire* pour leur réutilisation par les autres participants.

La Figure 82 présente un extrait en UML du vocabulaire initialisé par le chercheur intéressé par le sujet des chantiers, à l'aide de notre *processus de construction d'un vocabulaire formel*. Nous signalons également si l'élément a été trouvé dans le catalogue construit à partir d'OSMonto-fr et sa source.

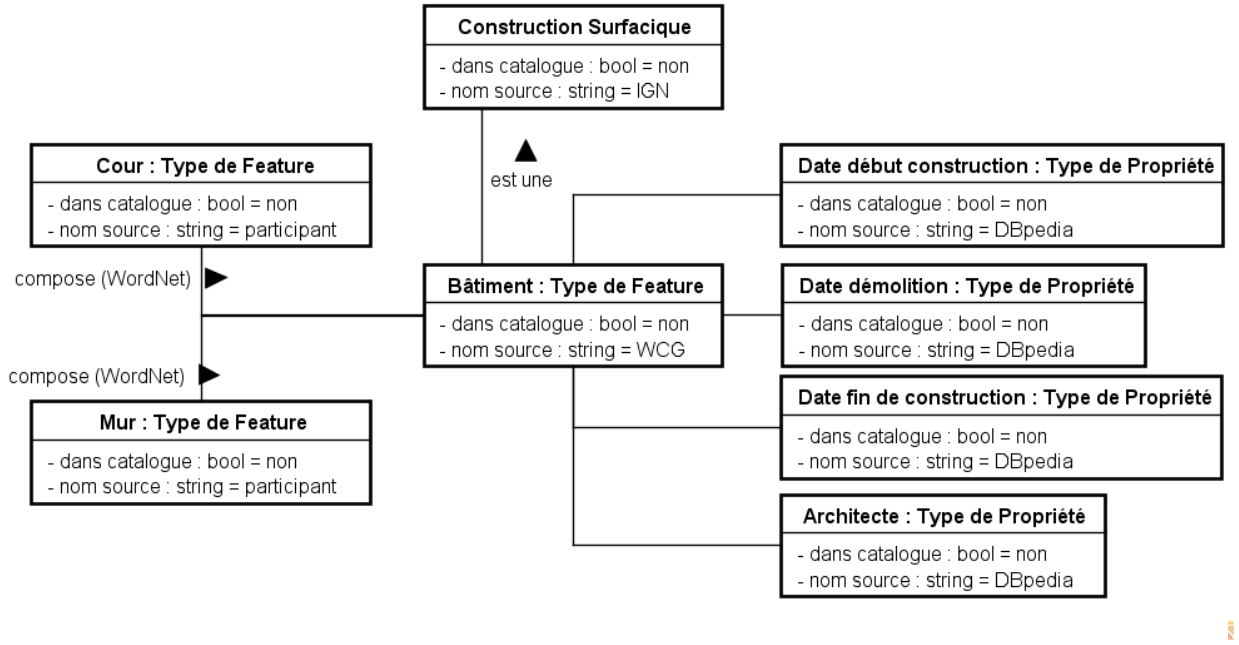



**Figure 82 : extrait de certains éléments de vocabulaire intéressant par le chercheur travaillant sur le thème des chantiers**

Le participant s'intéressait à décrire dans son vocabulaire, des éléments spécifiques sur les chantiers et le bruit qu'ils produisent. Sur ce dernier aspect, le processus n'était pas capable d'extraire des éléments satisfaisants pour le participant. Le participant a nommé `est une` le type de relation entre `Construction Surfacique` de l'IGN et ses `Bâtiments` identifié par le processus. Les types de relations trouvés grâce à WordNet permettent de détailler la représentation de la géométrie du bâtiment. Nous observons également des types de propriétés très pertinents pour le thème des chantiers. Nous avons relancé la recherche à partir des mots-clés, afin de trouver de nouvelles correspondances qui n'étaient pas dans le catalogue. En effet, nous avons trouvé un type de relation WordNet « `un trottoir peut composer une rue` ». Cet élément n'avait pas été instancié pendant l'initialisation à partir d'OSMonto-fr car un type de feature `trottoir` n'avait pas été défini. Dans ce cas, le participant signale son intérêt sur ce type de relation et ainsi nous initialisons manuellement un type de feature `Trottoir` de la manière suivante :

```
<featureType>
      <nom> Trottoir </nom>
      <instanciePar> Participant#3 </instanciePar>
      <source>
            <nom> Contributeur Particpant#3 </nom>
      </source>
</featureType>
```

### Cas #4 : un vocabulaire sur le tourisme

Le participant s'intéresse à construire un contenu collaboratif sur le thème du tourisme rural. Le Tableau 23 montre les mots-clés employés par le participant, le nombre d'éléments de vocabulaire pertinents selon le participant et le nombre total d'éléments trouvés par la méthode. Des nouveaux mots-clés ont été suggérés par le participant pendant l'expérience, c'est le cas de `sentier de randonnée`.

| Mot-clé | Nombre total d'éléments de vocabulaire trouvés par le processus | | Nombre d'éléments de vocabulaire choisi par le participant |
|---|---|---|---|
| | Extraits à partir du catalogue | Extraits à partir des bases d'extraction | |
| Hôtel | 34 | 0 | 14 |
| Sentier de randonnée | 17 | 1 | 11 |
| Route | 24 | 77 | 11 |
| Châteaux | 26 | 1 | 9 |
| Musées | 30 | 1 | 9 |
| Forêts | 16 | 27 | 6 |
| Cours d'eau | 0 | 25 | 5 |
| Patrimoine | 0 | 23 | 4 |
| Villes | 0 | 92 | 3 |
| Camping | 0 | 3 | 2 |
| Offices de tourisme | 1 | 0 | 1 |
| Réseau hydrographique | 0 | 0 | 0 |
| Chambre d'hôtes | 0 | 0 | 0 |
| Pente | 0 | 0 | 0 |
| **Total** | 148 | 250 | 75 |

**Tableau 23 : par mot-clé, le nombre d'éléments pertinents concernant le tourisme rural parmi le total d'éléments obtenus par le processus**

Nous pouvons remarquer la présence très importante d'éléments de vocabulaire concernant le thème du tourisme dans le *catalogue d'éléments de vocabulaire*, de même que dans les bases d'extraction. En effet, l'ontologie OSMonto-fr et les sources externes de vocabulaire utilisées par le processus semblent être très ciblées sur ce thème. Quelques exceptions apparaissent

sur les réponses vides obtenues à partir des mots-clés `réseau hydrographique`, `chambre d'hôtes` et `pente`. Les 75 éléments choisis par le participant sont copiés dans le *catalogue d'éléments de vocabulaire* pour leur réutilisation par les autres participants.

La Figure 83 présente un extrait en UML du vocabulaire initialisé par le chercheur intéressé par le thème sur le tourisme rural obtenu à l'aide de notre *processus de construction d'un vocabulaire formel*. Nous signalons également le cas où un élément a été trouvé dans le catalogue construit à partir d'OSMonto-fr et sa source.

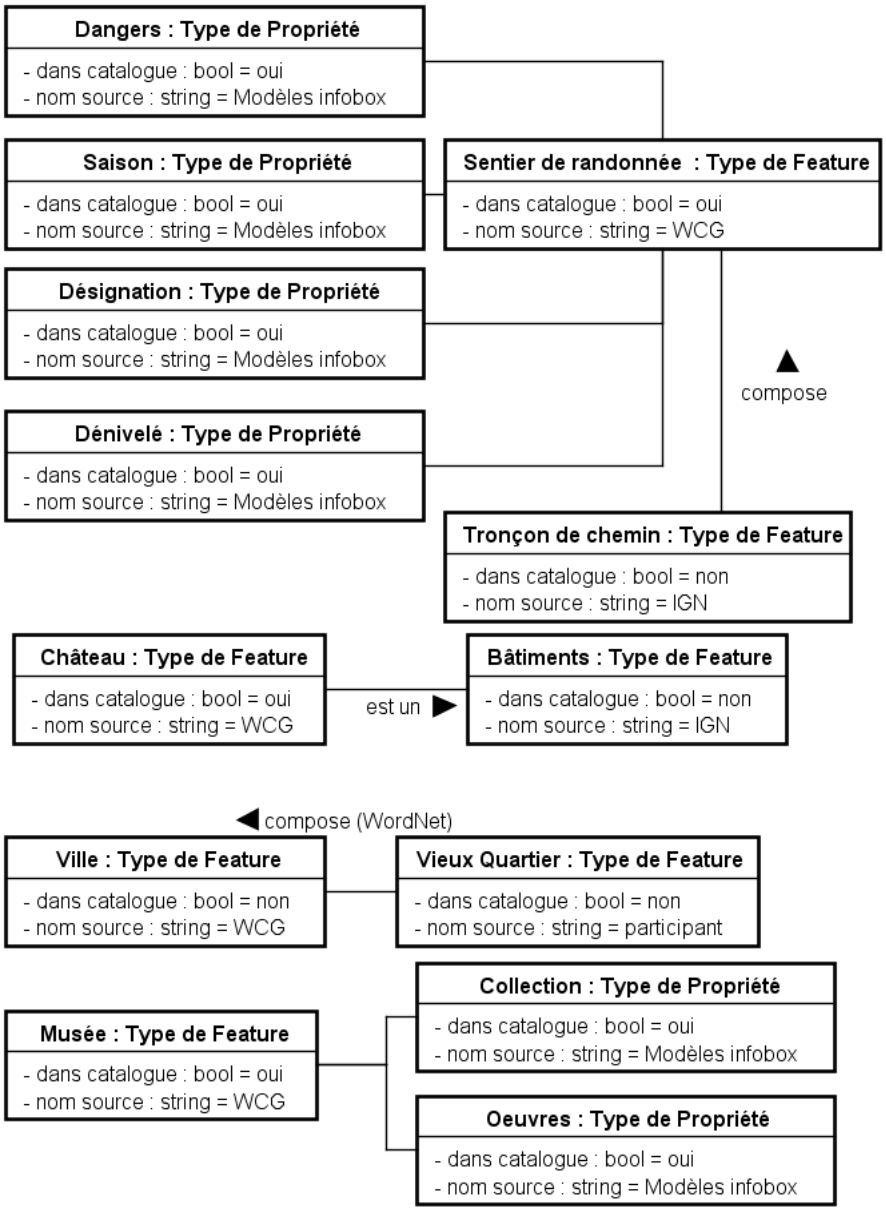



**Figure 83 : extrait de certains éléments de vocabulaire intéressants le chercheur dont les travaux portent sur le thème du tourisme rural**

Le participant s'intéressait à décrire dans son vocabulaire des éléments spécifiques sur les activités récréatives disponibles pour un touriste visitant un nouveau village. Le participant a apprécié la richesse au niveau des types de propriétés dans un contexte de tourisme comme ceux des sentiers de randonnée. Il a nommé un type de relation `est un` entre les châteaux et les bâtiments IGN repéré par le processus. De la même manière, le participant a signalé comme important des types de relations entre les types de *features* `Musée` et `Points d'intérêt et d'activité` de l'IGN, ainsi que le type de relation entre les types de *features* `Sentier de randonnée` et `Tronçon de chemin` du référentiel IGN. Un type de relation `compose` a été trouvé grâce à WordNet, entre le type de feature `Ville` (obtenu grâce au WCG) et un nouveau type de *feature* `Vieux quartiers`, une relation particulièrement pertinente dans un contexte touristique. Il faut noter que l'élément de vocabulaire `Bâtiment`, sélectionné pour ce vocabulaire sur le tourisme rural, n'est pas le même que l'élément du vocabulaire sur les chantiers en ville. Dans le premier cas, les bâtiments correspondent à ceux des spécifications IGN et seront utilisés comme référence. En revanche, les éléments de vocabulaire concernant les forêts introduits dans le catalogue d'éléments de vocabulaire par le participant intéressé aux déplacements de faune, étaient également utiles (et donc réutilisés) pour celui qui s'intéresse au tourisme.

### 2.2.3 Discussion sur l'expérience

Ces expériences avec les participants nous ont permis d'illustrer comment notre processus d'aide à la construction d'un vocabulaire fonctionne. Ces expériences nous ont également permis d'élargir la couverture de notre *catalogue d'éléments de vocabulaire* avec des nouveaux thèmes. Nous avons pu obtenir des remarques sur la perception des participants concernant une interface graphique conviviale pour notre processus. Par exemple, une interface dynamique de type « déposer et glisser « (*drag & drop*) où les éléments de vocabulaires sont des boîtes ou des bulles qui sont rapidement glissés de deux éléments d'interface « bases d'extraction » ou « catalogue » et déposées dans un élément d'interface « vocabulaire ». Un participant a particulièrement signalé le besoin de pouvoir « fusionner » deux types de propriétés trouvés par la méthode dans un nouveau type de propriété, et garder les deux liens de provenance.

Le fait de concevoir un processus afin d'aider une personne à enrichir au fur et au mesure un vocabulaire, comme nous l'avons fait, donne plus de flexibilité et de liberté à nos utilisateurs, un aspect essentiel dans un contexte collaboratif. En revanche, le fait de construire des vocabulaires statiques qui n'évoluent pas (ou très lentement) et qui ne reflètent pas les besoins changeants des utilisateurs peut décourager l'utilisation d'un système collaboratif. De plus, le fait d'utiliser des vocabulaires externes a permis d'avoir des éléments que les participants n'attendaient pas et qui étaient de leur intérêt, par exemple le type de propriété `nom en dialecte local` pour les villes, obtenus grâce à Wikipédia. En revanche, nous avons bien remarqué les faiblesses du processus : le nombre élevé de réponses renvoyées quand le processus accède aux bases d'extraction (et non le catalogue). Une autre faiblesse liée à la nature syntaxique des recherches effectuées par le processus, est la présence des problèmes

d'homonymies ainsi que l'absence de réponses. Nous avons également observé que dans plusieurs thèmes très « spécialisés », les vocabulaires disponibles ne suffissent pas. Cepedent, certains domaines sont couverts à un faible niveau de détail, par exemple, les batiments et leurs parties (ex : murs, porte, escalier, hall) décrits dans WordNet, et d'autres pas du tout, c'est le cas d'éléments de vocabulaire concernant par exemple, le bruit dans un chantier.

Ici, nous avons présenté les aspects de notre approche concernant la gestion de la cohérence au niveau du modèle afin de garantir l'homogéinité de la répresentation. Ci-après, nous nous concentrons sur la gestion de la cohérence des données.

# 3 Intégration des contributions dans un contenu collaboratif : la correction d'incohérences

Dans la stratégie d'intégration de contributions décrite en section III.3.2, le contributeur reçoit, selon l'incohérence, des suggestions sur des nouvelles géométries calculées à partir des méthodes correctrices. Elles sont implémentées dans un *catalogue de méthodes correctrices* indexé par contrainte. Ainsi, il peut choisir la meilleure méthode de correction entre plusieurs méthodes disponibles dans le catalogue. Cette section présente dans un premier temps la mise en œuvre dans le *catalogue de méthodes correctrices* d'une de ces méthodes et dans un second temps, son évaluation sur des incohérences issues des données OpenStreetMap.

## 3.1 Le catalogue des méthodes correctrices

Pour corriger une incohérence, des méthodes correctrices sont disponibles sous la forme d'un *catalogue de méthodes correctrices* indexé par les contraintes définies dans le système. Nous avons associé des méthodes correctrices à des contraintes afin de fournir au contributeur, pour une contrainte donnée, plusieurs possibilités de correction de l'incohérence associée. Par exemple, la méthode `shift` a été associée à la contrainte `les abribus ne chevauchent pas les routes de l'IGN`. Le pseudo-code de la méthode `shift` est montré ci-dessous et décrit ensuite.

```
1:  topologicalMap <- buildTopologicalMap(busShelterGeom, roadGeom)
2:  for all face in topologicalMap.faces do
3:    smallestPolygonPart <- chooseSmallestPolygon (face)
4:  end for
5:  lineOrientationAngle <-
6:    computeLineAbsoluteOrientation (roadGeom)
7:  translationDistance <-
8:    getSideLongMaxDistance (smallestPolygonPart, lineOrientationAngle)
9:  polygonOrientationAngle <-
10:   computePolygonAbsoluteOrientation (busShelterGeom)
11: rotationAngle <-
```

```
12:   computeAngleDifference (polygonOrientationAngle, lineOrientationAngle)
13: if busShelterGeom parallel to roadGeom
14:   newBushelterGeom <- translate (busShelterGeom, translationDistance)
15: else
16:   newBushelterGeom <- translate (busShelterGeom, translationDistance)
16:   newBushelterGeom <- rotate (newBushelterGeom, rotationAngle)
17: end if
18: return newBushelterGeom
```

La méthode `shift` construit en ligne 1 une carte topologique qui permet d'importer des objets (polygones, linéaires, et ponctuels) et de les traiter comme des objets topologiques (nœuds, arêtes, ou faces) (Grosso et al. 2012). Cette structure est utilisée dans les lignes 2-4 afin de pouvoir calculer l'ensemble des sous-polygones possibles obtenus par la coupure de la géométrie polygonale de l'abribus par la route. En lignes 5-6, l'angle d'orientation absolue du polygone est mesuré grâce à des opérations d'analyse spatiale (Duchêne et al. 2003). En lignes 7-8, la distance de translation, à laquelle le nouveau polygone sera déplacé est calculée c'est : la plus grande est choisie parmi toutes les distances entre un point du plus petit polygone à la route (*slandt distance*). En ligne 9-10, l'angle d'orientation absolue du polygone est calculé. Ensuite, il est possible, en lignes 11-12, de calculer la différence entre les angles d'orientation du polygone et de la ligne, afin d'obtenir l'angle duquel le nouveau polygone subira une rotation (angle de rotation). Si le polygone est parallèle à la ligne, une nouvelle géométrie est obtenue à partir de l'ancienne géométrie de l'abribus qui est déplacée selon la distance de translation calculée (voir lignes 13-15). Sinon, une nouvelle géométrie est également obtenue à partir de l'ancienne géométrie de l'abribus qui est déplacé selon la distance de translation et qui sera tournée de l'angle calculé.

## 3.2 Évaluation des méthodes correctrices

Ici, nous évaluons la méthode correctrice `shift` à partir des incohérences sur des données OSM.

### 3.2.1 Les incohérences sur des données OSM

Dans le catalogue de méthode correctrices, la méthode `shift` a été associée à une nouvelle contrainte « les bâtiments ne chevauchent pas les routes ». En effet, il est possible d'extraire des incohérences concernées par cette contrainte, identifiées grâce à l'outil Osmose[70]. Cet outil en ligne permet la recherche et la visualisation des incohérences dans OSM par région en France et par type (voir la section I.3.4). La Figure 84 montre un exemple d'une incohérence dans OSM de type « les bâtiments ne chevauchent pas les routes », signalé par Osmose. Il est valable d'associer cette contrainte à la méthode `shift` dans le *catalogue de méthodes*

[70] http://osmose.openstreetmap.fr/

*correctrices* car les géométries des bâtiments et des routes dans OSM possèdent des représentations polygonales et linéaires.

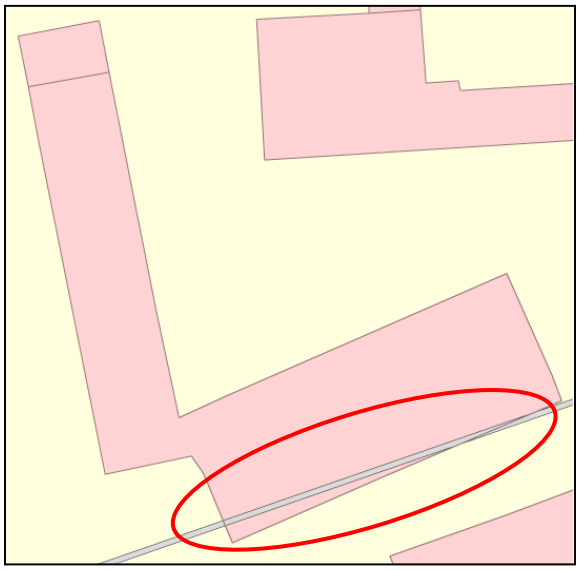


**Figure 84 : exemple d'une incohérence sur OSM de superposition entre un bâtiment et une route obtenus grâce à Osmose et visualisé sur l'outil d'édition Potlatch**

Il est possible de récupérer à partir de l'outil Osmose, pour une région donnée, une liste qui pointe vers les identifiants des bâtiments et des routes concernés par des superpositions. Plus précisément, nous avons récupérés la liste des 401 incohérences existantes à la date du 1 septembre 2012 sur la région française de Basse-Normandie[71] (voir la Figure 85). Chaque item de la liste correspond à des coordonnées géographiques en WGS84 positionnant l'incohérence, les identifiants de deux éléments et une description en langage naturel.

| pos | elems | subtitle |
|---|---|---|
| -1.30 48.73 | w 162123446 w 178847759 | Way intersecting building |
| -1.85 49.64 | w 153752597 w 178025616 | Way intersecting building |
| -1.85 49.64 | w 153752035 w 178025616 | Way intersecting building |
| -1.30 48.58 | w 162868836 w 177298193 | Way intersecting building |
| -1.30 48.58 | w 162869667 w 177298193 | Way intersecting building |
| -1.30 48.58 | w 162870144 w 177298193 | Way intersecting building |
| -1.30 48.58 | w 162869936 w 177298193 | Way intersecting building |
| -1.78 49.55 | w 100555059 w 177201771 | Way intersecting building |

…

**Figure 85 : extrait de la liste d'incohérences datées du 1 septembre 2012 portant sur la superposition entre un bâtiment et une route sur la région française de Basse-Normandie**

[71] http://osmose.openstreetmap.fr/utils/info.py?country=france_basse_normandie&item=1070

Une fois les incohérences identifiées, il est nécessaire d'obtenir les données OSM correspondant, respectivement au réseau routier et bâtiments sur la Basse-Normandie (voir la Figure 86), les jeux de données ont été obtenus à partir du site de Planet OSM[72].

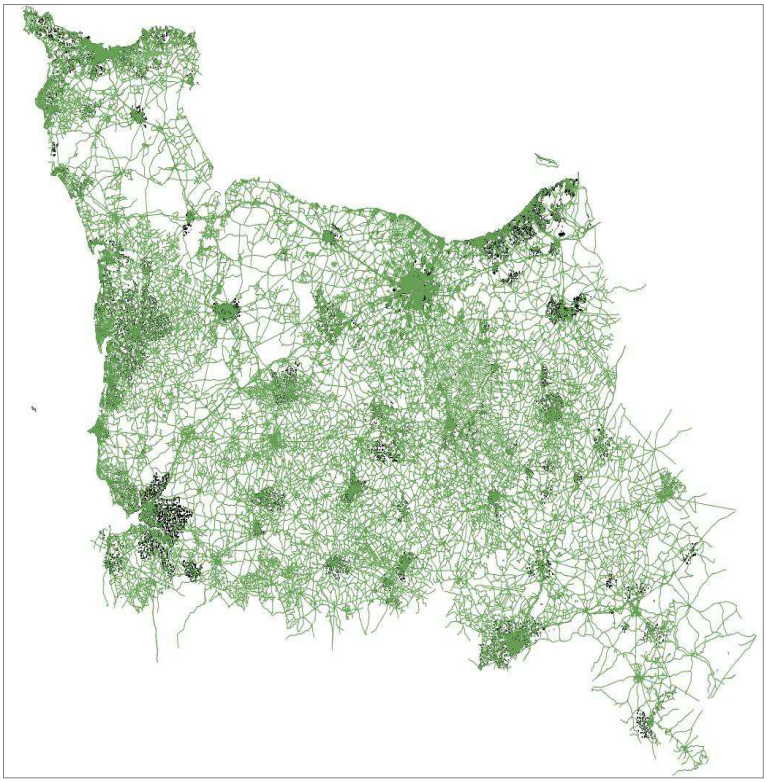


**Figure 86 : données OSM datant du 3 septembre 2012 existantes sur la région Basse-Normandie correspondants aux thèmes routes (en vert) et bâtiments (en noir)**

Une fois les données chargées et les incohérences identifiées, il est possible de les visualiser (voir la Figure 87). Nous observons en détail les caractéristiques suivantes : la complexité de l'intersection entre le bâtiment et la route, le nombre de routes intersectant le bâtiment, les valeurs des tags des objets impliqués, la nature des objets autour des objets concernés par ces incohérences.

[72] http://planet.osm.org/

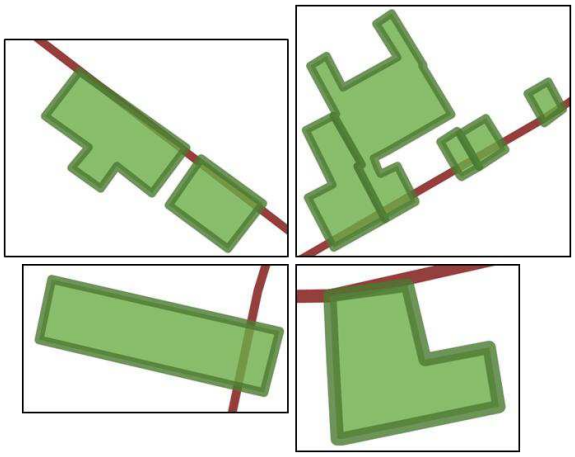


**Figure 87 : des incohérences entre des bâtiments et des routes OSM**

Nous observons souvent que ces incohérences sont liées à l'import automatique du cadastre français, nous les appelons des « incohérences en lot ». La Figure 88 montre deux bons exemples de géométries de bâtiments provenant du cadastre français (données fournies par la Direction Général des Impôts), qui ont été importées dans OSM. En effet, il est possible de vérifier la source des données (`source = cadastre-dgi-fr source : Direction Générale des Impôts – Cadastre Mise à jour : 2011`). Il est aussi possible de constater l'import automatique car il a été effectué par des contributeurs possédant des comptes appelés *bots*, dédiés à des tâches automatiques.

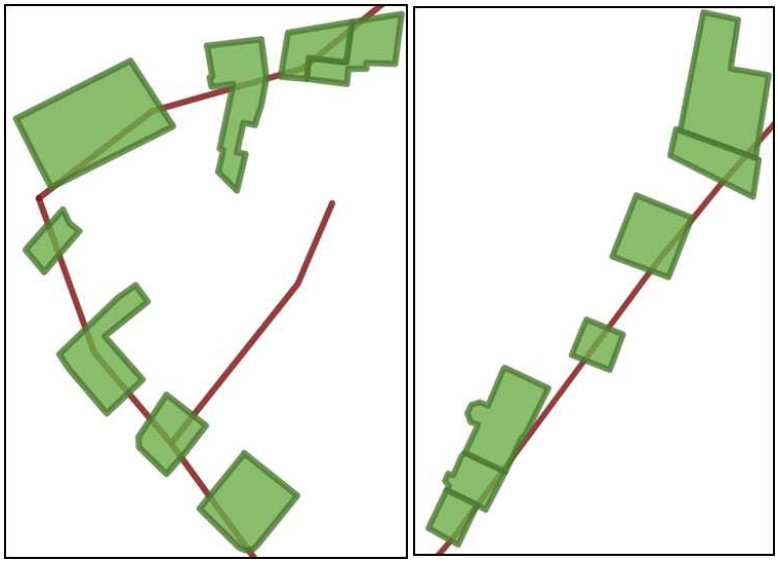


**Figure 88 : incohérences en lot provenant de l'import automatique du cadastre dans OSM**

Une autre incohérence trouvée moins souvent concerne des intersections « complexes » entre des bâtiments et des routes. Notre méthode `shift` assume que le bâtiment est découpé par l'intersection d'uniquement une route, c'est-à-dire le nombre de routes intersectant le bâtiment

est supérieur à un. Cependant, cela n'est pas toujours le cas, une evidence de ce type d'incohérence est présentée sur la Figure 89.

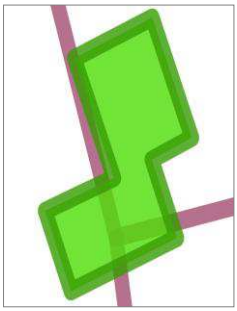


**Figure 89 : un exemple d'une incohérence sur OSM concernant un bâtiment chevauchant une route qui n'a pas été corrigé par notre méthode**

Nous observons également que certaines incohéreces sont liées au mauvais étiquetage (*tagging*) des objets. En effet, nous remarquons une quantité considérable d'objets incorrectement étiquetés comme des bâtiments. La Figure 90 illustre des incohérences identifiées comme des «  bâtiments » (`building = Yes`) et des routes. Ces objets sont en réalité des îlots de bâtiments, ou dans un cas en particulier, une muraille entourant le domaine d'un château.

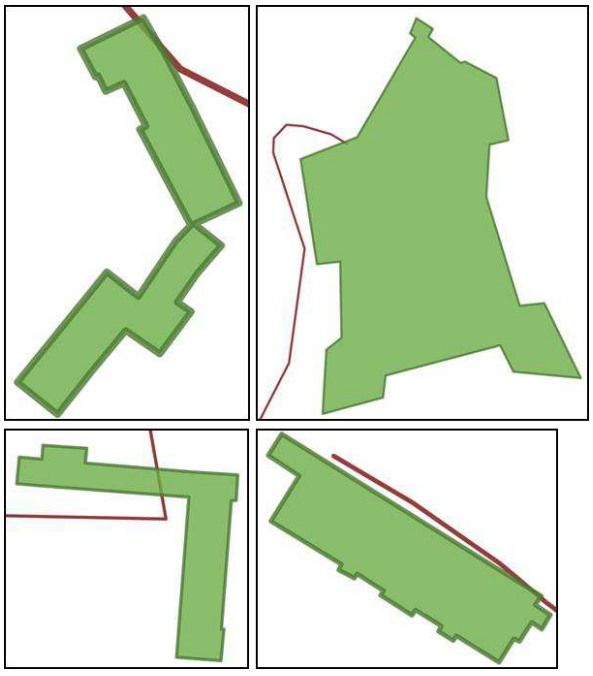


**Figure 90 : mauvais étiquetage des objets polygonales (en vert) étiquetés comme des bâtiments (`building = Yes`)**

Une derniere incohérence que nous trouvons moins fréquemment est celui lié à la mauvaise interprétation des spécifications OSM par un contributeur. Par exemple, nous remarquons un contributeur qui a dessiné l'emprise d'un pont et celle-ci a identifiée comme un bâtiment (`building = Yes`). La representation de cet objet en format OSM XML est présentée ci-dessous et il peut être visualisé sur la Figure 91.

```
<osm version="0.6">
<way id="157546464" visible="true" timestamp="2012-08-06T21:05:36Z"
version="3" changeset="12638992" user="mobip" uid="2983">
      <nd ref="1697800459"/>
      <nd ref="1725649107"/>
      …
      <tag k="building" v="yes"/>
      </way>
</osm>
```

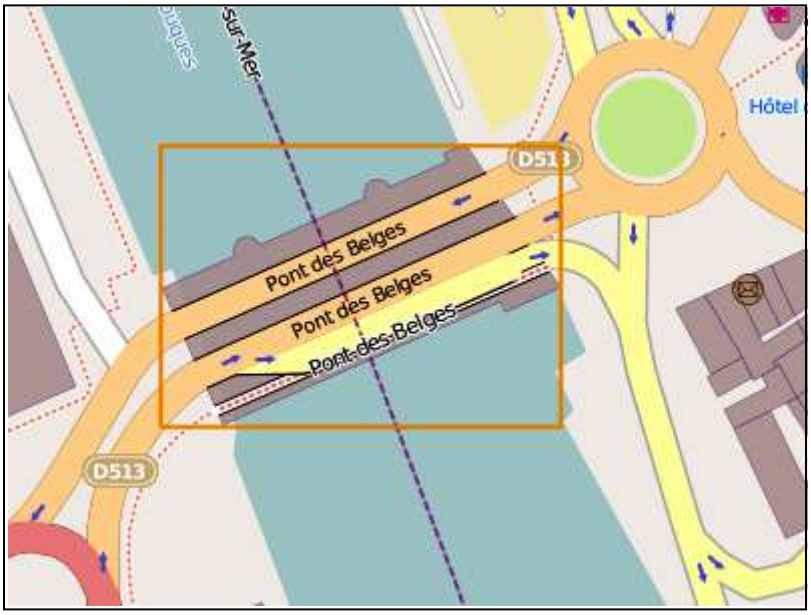



**Figure 91 : mauvaise interprétation des spécifications : l'emprise d'un pont a été dessinée et identifiée comme un bâtiment (`building = Yes`)**

Ce contributeur dessine l'emprise du pont en sachant que la géométrie du pont a été déjà representée comme un tronçon de route. En regardant l'historique des données impliquées dans cette incohérence, nous constatons qu'il a été ajouté il y a des mois par un contributeur, ensuite mis à jour deux fois par le même contributeur et puis par un autre. Malheureusement, le tag source est vide. Nous constatons de cette façon que c'est une erreur introduite par un humain. En effet, les commentaires des groupes de modification correspondants à toutes les

versions de l'objet viennent d'un humain[73], et non d'un contributeur de type *bot*. Il est possible que l'intention du contributeur fût de representer un bâtiment qui est utilisé comme un pont, comme par exemple le Ponte Vecchio de la ville italienne de Florence. Cependant, il existe un tag `building = bridge` pour ce propos.

### 3.2.2 Correction d'incohérences

Nous appliquons la méthode de correction `shift` pour les 401 incohérences obtenues grâce à Osmose et nous constatons que 214 incohérences n'ont pas été correctement corrigées. Les distances de déplacement des bâtiments calculés par la méthode n'ont pas été suffisamment larges pour corriger ces incohérences. Une deuxième évaluation de la méthode sur ces 214 incohérences, en considérant les nouvelles géométries des bâtiments proposées auparavant par la méthode, permet de corriger 109 incohérences. Ainsi, un total de 105 incohérences de 401, environ 25%, n'ont pas été automatiquement corrigés après vérification visuelle car l'intersection reste toujours entre le bâtiment et la route. Par exemple, la Figure 92 montre une incohérence comme celle décrite précédemment.

Ensuite, nous vérifions les nouvelles géométries proposées par la méthode pour les 296 incohérences corrigées. Nous constatons que les corrections des 201 incohérences restent cohérentes par rapport à d'autres données autour de ces incohérences. Certaines de ces corrections peuvent être visualisées dans la Figure 92. Nous remarquons également que la méthode est capable de corriger des incohérences en lot liés à l'import automatique du cadastre français dans OSM, comme ceux présentés en Figure 93

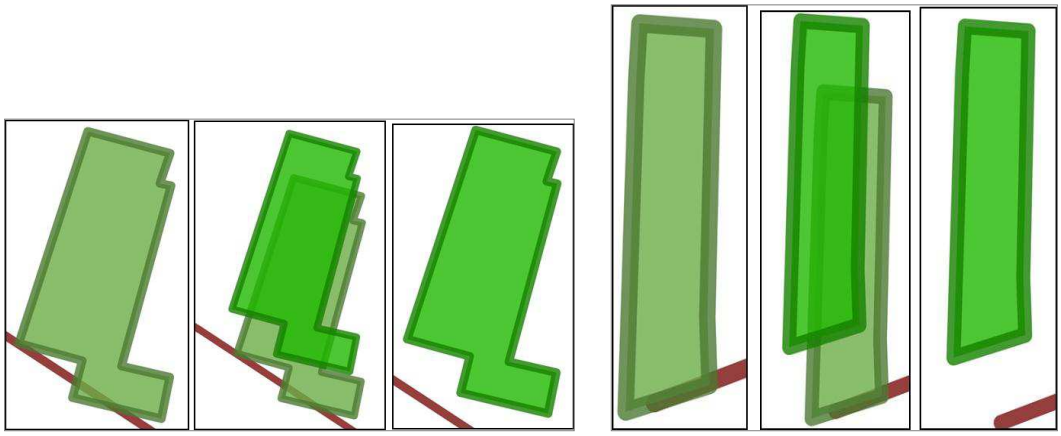


**Figure 92 : chaque image montre une incohérence de superposition d'un bâtiment et une route dans OSM : (à g.) incohérence détectée, (au c.) incohérence détectée et correction proposée, (à d.) correction acceptée**

[73] Il est aussi possible de connaître qui a fait quoi dans OSM : http://simon04.dev.openstreetmap.org/whodidit/?zoom=11&lat=47.97428&lon=-3.46791&layers=BTT&age=1%20month

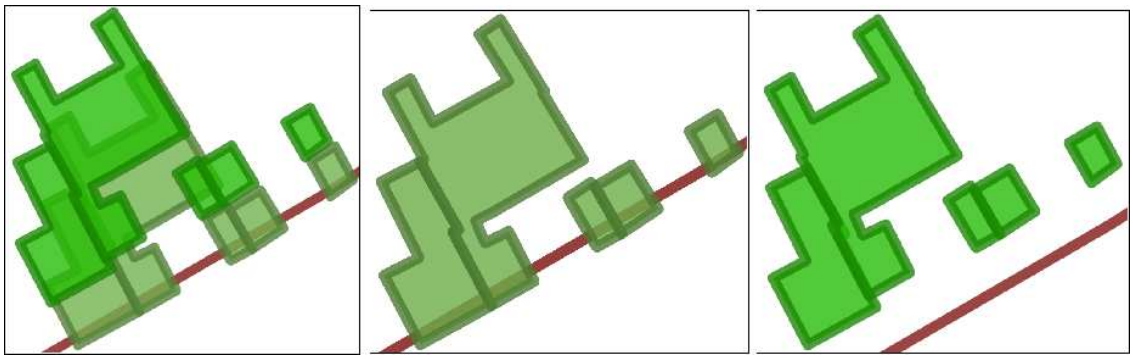


**Figure 93 : incohérence en lot provenant de l'import automatique du cadastre dans OSM**

### 3.2.3 Discussion

Cette évaluation nous a permis d'ajuster les paramètres de notre méthode de correction afin d'améliorer la précision de la correction. Concernant les incohérences qui ne sont pas corrigeables par la méthode, nous envisageons quelques solutions qui peuvent aider à améliorer cette méthode. Pour alerter le mauvais étiquetage d'un objet à un contributeur, il est possible de vérifier la superficie du polygone afin de déterminer si elle est trop grande par rapport à des bâtiments en considérant si la zone est rurale ou urbaine. De plus, il faut absolument considérer les objets autour des incohérences afin de traiter correctement des incohérences comme celle de la figure 87. En effet, notre méthode déplace bien l'emprise du pont loin du tronçon de route mais l'objet est positionné sur le fleuve. Néanmoins, la méthode ne corrige pas la nouvelle incohérence concernant le flottement de l'emprise du pont sur la surface d'eau. Cette expérimentation illustre donc le besoin de définir des nouvelles contraintes afin de traiter automatiquement plus d'incohérences.

Il est sans doute souhaitable d'intégrer des nouvelles méthodes pour la correction d'incohérences afin de pouvoir proposer plusieurs corrections possibles pour une même incohérence (ex : le bâtiment qui chevauche la route). Ces méthodes peuvent s'appuyer sur des librairies d'analyse spatiale utilisées par les processus automatiques de généralisation cartographique. La plate-forme CartAGen du laboratoire COGIT propose des algorithmes pour détecter par exemple, la direction d'échappement d'un bâtiment coincé[74], l'alignement d'un groupe de bâtiments afin de les déplacer en bloc[75], et le déplacement de bâtiments tenant en compte de bâtiment autour (Ruas 1999; Renard et al. 2011; Duchêne et al. 2012).

[74] http://www.openstreetmap.org/?box=yes&bbox=-1.58931%2C48.8342%2C-1.58923%2C48.83426
[75] http://www.openstreetmap.org/?lat=49.181095&lon=-1.574108&zoom=18

# Bilan et Conclusion

La prolifération de projets communautaires de cartographie collaborative met en évidence le besoin de faciliter, pour les individus, la constitution et l'édition des contenus géographiques. Ces contenus peuvent être particulièrement ciblés sur des thèmes divers comme le tourisme, la biodiversité, ou encore l'écologie. Le caractère collaboratif de ce mode de production impacte la cohérence de ces contenus. Au cours de ce mémoire thèse, nous avons souligné l'importance de concevoir des systèmes collaboratifs qui aident les individus à constituer et à gérer la cohérence de leurs contenus afin de permettre que des prises de décision s'appuient dessus. Nous avons proposé un modèle, baptisé *Coalla*, pour l'édition collaborative et la gestion de la cohérence d'un contenu géographique.

L'édition collaborative d'un contenu géographique renvoie à deux contextes de production, celui des producteurs institutionnels de données géographiques comme l'IGN et celui des projets communautaires de cartographie collaborative comme OpenStreetMap. Pour mieux décrire et comprendre ces contextes de production, nous avons participé à des activités de collecte ainsi que de saisie de données organisées par l'IGN, de même que par OSM. *A priori,* ces deux contextes semblent de nature très différente. Néanmoins, nous avons constaté globalement plus de ressemblances que de divergences vis-à-vis des pratiques suivies par les opérateurs/contributeurs et en regard de la gestion des données. Une différence importante entre ces deux contextes est l'exhaustivité des spécifications, ainsi que l'homogénéité de couverture du territoire pendant l'acquisition des données. En effet, les opérateurs d'un producteur national collectent les données en suivant les spécifications d'acquisition de manière homogène dans le territoire. Une autre différence est l'assiduité en regard des évaluations de qualité et leur documentation. En analysant ces contextes de production, il a été possible d'identifier les problèmes de recherche à aborder dans la production collaborative de données géographiques.

Une caractéristique de notre approche est la prise en compte de plusieurs domaines de recherche qui sont capables d'apporter des solutions. Ainsi, nous avons investigué des domaines de la littérature scientifique assez vastes afin d'identifier certains mécanismes qui semble pertinents pour la gestion de la cohérence d'un contenu géographique collaboratif. Particulièrement, nous avons considéré des expertises en information géographique de même que des atouts du Web collaboratif et du libre. Les contributions de ce travail de thèse, présentés en chapitre III, sont résumées ci-après.

La **première contribution**, présentée dans la section III.1, est l'identification des éléments qui semblent pertinents pour faciliter la gestion de la cohérence d'un contenu géographique. Cette notion de cohérence est directement liée à l'homogénéité de la représentation et au respect des relations spatiales importantes. Plus précisément, nous avons identifié les éléments que doit

comporter un vocabulaire partagé par les contributeurs visant à faciliter la construction et la gestion de la cohérence d'un contenu géographique collaboratif. Ces éléments servent à définir des relations importantes, souvent implicites et retrouvées grâce aux géométries, qui doivent être préservées dans le contenu. Dans notre modèle, des types de relations au niveau du schéma de données pourront être associées à des mécanismes d'évaluation pour vérifier automatiquement que les relations sont préservées. Ces types de relations sont définis grâce à des contraintes d'intégrité que le contributeur veut voir respectées dans son contenu. Ces contraintes peuvent être définies entre le contenu collaboratif et un contenu de référence de l'IGN, au niveau du schéma ou au niveau des instances. Nous proposons de permettre à un contributeur d'expliciter des relations entre le contenu collaboratif et le contenu de référence afin de faire l'ancrage entre ces deux contenus. Cela permet à des processus automatiques de gérer la cohérence du contenu. Les contributeurs d'OSM confirment dans les résultats de notre sondage, décrits en Annexe A, l'importance d'expliciter des propriétés et des relations spatiales, et de les préserver dans le contenu car elles sont utiles pour les applications. En effet, certains contributeurs se sont exprimés en indiquant que la saisie des relations comme « le bâtiment en face de l'église et à droite du bureau de la Poste », peut améliorer l'utilisabilité du contenu pour la navigation et le repérage.

La **deuxième contribution**, présentée dans la section III.2, est un processus pour aider les contributeurs à construire à la volée, un vocabulaire formel en lui suggérant des éléments de vocabulaire pour la gestion de la cohérence et l'utilisabilité du contenu. Grâce au sondage proposé aux contributeurs OpenStreetMap décrit en Annexe A, nous avons constaté qu'une partie considérable de ces contributeurs préfèrent manipuler les géométries des entités au lieu d'expliciter des propriétés et des relations spatiales. Cela nous a permis de confirmer notre intuition sur le besoin d'acquérir ces informations, les définir, et les intégrer dans un vocabulaire formel. Le système encourage la réutilisation de ce vocabulaire grâce à un catalogue d'éléments de vocabulaire, afin d'aider les contributeurs à se mettre en accord sur la modélisation de la partie du monde qu'ils veulent représenter. Notre processus s'appuie sur des sources diverses. D'un côté, l'utilisation des sources communautaires comme Wikipédia et DBpedia permet l'incorporation d'un vocabulaire dont les termes sont fréquemment utilisés par les communautés du libre. Nous avons décidé de conserver les deux sources Wikipédia et DBpedia car l'encyclopédie reste toujours la plus à jour des deux. De l'autre côté, l'utilisation des sources issues de professionnels permet de s'appuyer sur des modèles formels reflétant les points de vue de ces organisations. En particulier, il est important de créer des références avec des données d'autorité comme celles de l'IGN qui peuvent servir de référentiel. Ce processus peut être utile pour l'aide à la construction d'un contenu thématique ciblé, par exemple zone humides, relevés faunistiques et floristique, en montrant à un contributeur ces deux points de vue, communautaire et professionnel. Cela est aussi valide dans le cas d'OSM. Nous pouvons prendre l'exemple d'un contributeur qui s'est exprimé dans notre sondage de la manière suivante : "*il faudrait offrir la possibilité de varier les thématiques et la classification des objets proposée par les tags est actuellement insuffisante*".

L'évaluation de cette contribution concernant la construction et la réutilisation du catalogue d'éléments de vocabulaire est présentée en chapitre IV. Nous avons initialisé notre catalogue

d'éléments de vocabulaire à partir des tags OSM grâce à l'ontologie OSMonto. Il a été également possible de comparer cette ontologie avec les différents vocabulaires utilisés par notre processus. Nous avons constaté les liens forts entre les communautés OSM et Wikipédia vis-à-vis de la couverture de thèmes, de même que les différences avec des vocabulaires experts. Ces analyses sur la couverture de thèmes d'un vocabulaire par rapport à un autre vocabulaire peuvent aider à découvrir comment l'un peut aider à enrichir l'autre. Par exemple, les vocabulaires collaboratifs pourraient bénéficier de l'expertise contenue dans les schémas et ontologies produits par les producteurs institutionnels. L'expérience concernant la construction d'un vocabulaire formel avec des chercheurs du Laboratoire COGIT nous a permis de découvrir les atouts et les faiblesses de notre prototype. Un atout est la possibilité d'offrir des suggestions très diverses concernant certaines thématiques typiquement "populaires" comme par exemple celle du tourisme. Une faiblesse importante est l'interface graphique proposée. En effet, elle doit être conviviale pour améliorer l'utilisabilité de l'application. Un bénéfice important de cette expérience a été la possibilité d'enrichir notre catalogue d'éléments de vocabulaire sur des thèmes comme les déplacements de faune et les littoraux.

La **troisième contribution**, présentée dans la section III.3, est une stratégie d'évaluation et de réconciliation des contributions afin de les intégrer d'une façon cohérente au contenu géographique collaboratif. Cette stratégie est basée sur l'évaluation des contraintes d'intégrité notamment sur des types de relations, et des contraintes de dépendance notamment sur les types de propriétés. Afin d'aider le contributeur à gérer la cohérence du contenu, la méthode inclut la notion de méthode correctrices afin de corriger les incohérences en suggérant une géométrie alternative. Afin de permettre l'édition concurrente du contenu, notre méthode effectue aussi la fusion de séquences d'éditions provenant des contributeurs différents et peut considérer l'existence d'un historique d'éditions pour réconcilier deux contributions en conflit. Pour la fusion de deux séquences d'édition, nous avons traité en premier lieu, les conflits de création d'objets, c'est-à-dire, la création de deux objets différents qui représentent la même entité du monde réel. Ce type de conflit peut arriver plus souvent dans un contexte de production communautaire où il existe des milliers de contributeurs qui travaillent à distance. Notre contribution pourrait être utile par exemple dans le contexte d'une cartopartie OSM. D'autre part, nous avons réalisé des entretiens avec des professionnels IGN et une comparaison de notre modèle avec celui de l'IGN concernant la gestion de l'historique et des différentiels de données. Cette information est présentée dans la section III.4. Nous avons constaté que l'existence d'un historique d'éditions peut favoriser davantage l'analyse des patterns d'éditions des opérateurs et l'édition concurrente de zones indépendantes du contenu. Ces aspects sont aussi importants dans le contexte des projets communautaires. L'évaluation de cette contribution concernant la correction automatique d'incohérences à partir d'un cas basé sur des données OSM, est présentée en chapitre IV. Notre méthode est capable de corriger automatiquement la plupart des incohérences d'un type choisi. Ce test nous a permis d'investiguer la nature de certaines incohérences concernant les *imports* automatiques dans OSM et leur impact sur la cohérence des données. Evidemment, il y un besoin d'incrémenter le nombre de méthodes pour la correction automatique d'incohérences, et possiblement de réaliser souvent ce types d'analyses et de les documenter.

# Perspectives

Les perspectives de recherche liées directement ou indirectement à l'édition collaborative et à la gestion d'un contenu géographique, sont prometteuses et nombreuses. Certaines d'elles sont décrites ci-dessous.

## La prise en compte des relations spatiales pour la cohérence

L'explicitation de relations spatiales pourrait constituer un nouveau mode de saisie pour des contributeurs novices. Quelques contributeurs se sont exprimés dans notre sondage en indiquant que la saisie des relations comme « le bâtiment en face de l'église et à droite du bureau de la Poste » peut constituer un nouveau mode de contribution dans OSM. Tout de même, la plupart de contributeurs qui ont participé dans notre sondage, ont exprimé leur réticence à expliciter des relations spatiales dans le contenu. Ainsi, il serait aussi intéressant de pouvoir les acquérir automatiquement en exploitant des nouvelles sources afin de proposer plus de suggestions au contributeur. Ces sources peuvent être d'origine communautaire, comme le graphe RDF OSM Semantic Network (Ballatore et al. 2012). En effet, ce graphe est généré automatiquement à partir du site Map Features d'OSM et reflète plus de relations implicites entre les tags décrits sur ce site. Des nouvelles sources de vocabulaire à incorporer également peuvent être des ressources linguistiques en français constituées par la communauté française en traitement automatique de langages naturels (TALN). De cette manière, il est possible de mieux traiter par exemple, les problèmes d'homonymies. Une approche formelle du raisonnement qualitatif spatial très compatible avec le langage naturel est celle du *Region connection calculi* (RCC). Cette formalisme s'intéresse à la représentation non numérique mais avec des symboles d'une configuration d'un espace à partir d'un ensemble "base" de relations spatiales (Renz et Nebel 1998). Ainsi, Il est possible de détecter des incohérences dans les descriptions données par les gens par des méthodes de raisonnement comme l'algorithme *path-consistency* (Egenhofer et Franzosa 1991). Particulièrement, Roussey et Pinet (2010) utilisent une approche basée sur la logique de description (DL) pour formaliser un ensemble de relations spatiales qualitatives afin qu'un moteur d'inférence puisse évaluer la cohérence d'un ensemble de relations spatiales.

## Les contributeurs et leurs contributions

L'approche sur la confiance et la réputation proposée par Bishr et Kuhn (2007) et Bishr et Janowicz (2010) serait intéressante pour notre stratégie de réconciliation dans un contexte communautaire. Il est ainsi possible de modéliser la confiance d'une contribution et la réputation d'un contributeur comme une note (*score*). La contribution qui possède la meilleure note et le contributeur avec la meilleure réputation seront prioritaires dans le choix de réconciliation.

Il serait intéressant d'étudier l'activité d'édition des contributeurs en effectuant des analyses quantitatives sur l'historique de données. Il serait possible de trouver des *patterns* de contributions fréquents, des "mauvaises habitudes" des contributeurs, ou des conflits d'édition fréquents. L'analyse des patterns de contribution et des conflits édition est également effectuée pour Wikipédia depuis quelques années (Kittur et al. 2007; Yasseri et al. 2012).

La définition d'une typologie d'incohérences est une idée intéressante que nous avons eue grâce au sondage. Il existe un intérêt récent sur ce sujet dans la communauté de géomatique intéressée par OSM. En effet, le projet OSM-Great Britain[76] s'intéresse à l'étude des incohérences dans OSM et leurs possibles corrections. Ils ébauchent une liste de possibles incohérences qui peuvent se passer dans OSM. En consensus avec le projet OSM-GB et la communauté de contributeurs, il serait intéressant de définir une typologie d'incohérences.

## L'intégration et les mises à jour

Il est important d'investiguer les besoins des contributeurs dans les projets communautaire. Le sondage proposé aux contributeurs d'OSM a mis en évidence des besoins qui pourraient être adressés dans des travaux ultérieurs : l'aide à l'intégration de données issues de l'*open data,* et la conception de nouveaux modes de saisie mieux adaptés aux contributeurs novices. Ce sondage que nous avons proposé est un premier pas intéressant dans l'évaluation des besoins des contributeurs.

De plus, l'exploitation de sources communautaires (ex : Wikipédia) serait intéressante pour mettre en place des systèmes automatiques d'alertes pour la mise à jour des bases de données géographiques, particulièrement celles mises à jour par des producteurs institutionnels comme l'IGN. Par exemple, un mécanisme automatique mis en place par l'IGN pourrait rechercher automatiquement des possibles changements soufferts par des entités géographiques et qui sont décrits dans Wikipédia. D'ailleurs, il ne faut pas se limiter à Wikipédia mais explorer d'autres sources collaboratives existantes sur le Web.

## L'interopérabilité et le choix d'architecture

Le suivi des normes de l'OGC est essentiel pour développer des applications interopérables manipulant des données géographiques. Ainsi, il existe une nouvelle proposition, soumise récemment au Consortium Open Geospatial (OGC), appelée *GeoSynchronization Services*[77]*.* Cette proposition vise à modéliser et à traiter les aspects de l'édition collaborative d'un contenu géographique selon les standards de l'OGC. En particulier, ils considèrent des rôles pour les contributeurs, les gens peuvent par exemple s'inscrire comme des lanceurs d'alerte pour signaler des erreurs ou comme des validateurs pour accepter les nouveaux changements du contenu.

[76] www.osmgb.org.uk

[77] http://www.opengeospatial.org/pressroom/pressreleases/1308

La mise en place d'une architecture décentralisée basée sur un réseau pair à pair est intéressante pour faciliter le travail déconnecté. L'inconvénient d'une approche centralisée comme la notre, est que les éditions des clients doivent arriver au serveur pour être réconciliées. Dans le cas de la perte du serveur, il n'est pas possible de traiter la demande du client. Ainsi, il existe un besoin d'approches décentralisées comme celles des éditeurs collaboratifs de documents XML et des moteurs de Wiki P2P (Weiss et al. 2007).

# Annexes

# Annexe A - Synthèse des réponses à un sondage proposé aux contributeurs d'OSM France

Ce document est une synthèse des réponses du sondage réalisé pour étudier le retour des contributeurs sur le projet de cartographie collaborative OSM. Ce sondage a été ouvert et envoyé sur la liste de diffusion OSM France le 31 janvier 2012, et clôturé le 15 février 2012. Ce sondage couvre globalement cinq aspects : la formation des contributeurs, les modes de contribution préférés, la satisfaction des contributeurs à des outils d'édition, les relations spatiales (pour améliorer l'utilisabilité et faire l'ancrage vers un référentiel de données) et les conflits expérimentés d'édition collaborative.

**1. Le sondage**

Les questions du sondage sont listées ci-dessous :

*Question#1. Avez-vous réalisé des études en informatique, géographie, cartographie ou géomatique ?*
*oui*
*non*

*Question#2. Quelles sont les modes de contribution sur OpenStreetMap que vous utilisez plus fréquemment ?*
*Trace des images satellites*
*Collecte de traces GNSS et téléchargement sur le serveur OpenStreetMap*
*Chargement de sources de données autoritaires disponibles (ex : cadastre, Corine Land Cover)*
*Renseignement des informations sémantiques (tags) en utilisant des applications sur smartphones (ex : Mapzen POI collector sur Iphone)*
*Correction d'erreurs sur des outils OSM (ex : Keep Right, OpenStreetBugs)*
*Autre (veuillez préciser)*

*Question#3. Êtes-vous satisfait de la façon dont vous pouvez saisir des données dans OpenStreetMap ou des outils d'édition collaborative similaires (ex : Wikimapia) ? Si vous pensez à quelque chose qui vous semble particulièrement utile ou quelque chose qui manque et serait apprécié, pouvez-vous le mentionner ?*

*Question#4. Pensez-vous que la description de caractéristiques de certains objets comme «c'est le plus grand immeuble de la zone » ou de relations entre certains objets comme « l'abribus est exactement en face du bureau de poste » peut rendre le contenu encore plus utile ? Si non, pourquoi ? Si oui, avez-vous des exemples en tête, en particulier des relations entre des nouveaux objets et des objets topographiques de référence comme ceux présents sur les cartes de sources gouvernementales comme l'IGN (routes, bâtiments, etc.) ?*

*Question#5. Avez-vous déjà connu un conflit d'édition entre contributeurs ? Si oui, pouvez-vous le décrire ?*

Les deux premières questions sont "fermées", c'est-à-dire, à choix simple et multiple. La Question#1 vise à connaître la proportion de contributeurs OSM qui ont reçu une formation professionnelle dans le domaine de l'informatique, la géomatique, la cartographie ou la géographie. La Question#2 porte sur les modes de contributions les plus utilisés par les contributeurs dans OSM.

Les trois dernières questions sont "ouvertes". Les réponses sont donc exprimées sous la forme de texte libre. Il existe plusieurs limitations à cette approche. Tout d'abord, il peut s'avérer difficile de comprendre le sens d'une réponse car le langage humain est en soi ambigu. De plus, nous ne pouvons pas faire d'analyse quantitative des réponses (ce qui est habituellement fait pour analyser les résultats d'un sondage). Néanmoins, le principal avantage du texte libre est qu'il permet de clarifier, comprendre et expliquer une réponse (Taylor-Powell et Renner 2003). Ainsi, la Question#3 cherche à savoir si les contributeurs sont satisfaits des outils d'édition et les modes de saisie. La Question#4 a l'intention de connaître la perception des contributeurs concernant l'utilisation des relations spatiales explicites afin d'améliorer l'utilisabilité du contenu et de faire un ancrage vers un référentiel de données. La Question#5 vise à identifier les conflits d'édition collaborative les plus fréquents qui existent entre les contributeurs.

## 2. La synthèse des réponses

Un total de cinquante-huit contributeurs ont rempli le sondage publié en ligne. Pour chaque question fermée, nous quantifions simplement le nombre de réponses différentes et présentons la réponse dans un tableau. Pour chaque question ouverte, nous effectuons une analyse exhaustive des réponses contenues (Huberman et al. 2003), cette méthode d'analyse qualitative des données est habituelle dans les sciences sociales. Plus précisément, il s'agit pour chaque question du sondage, de créer manuellement des catégories cohérentes pour rassembler les différents sujets traités par les participants. Nous fabriquons ainsi une carte conceptuelle, connue aussi sous le nom de *mind mapping*, celle-ci permet de résumer la structure synthétique d'une connaissance construite à partir de sources diverses (Carter et Frith 1998).

### 2.1 Première question : formation professionnelle des participants

Le but de cette question est de savoir si les participants ont été formés dans un domaine lié à l'informatique, la géographie, la cartographie ou la géomatique, c'est-à-dire, à des domaines liés à l'information géographique. Les réponses sont résumées dans le Tableau 24.

| Oui | Non |
|---|---|
| 32 (55,2 %) | 26 (44,8 %) |

**Tableau 24 : formation des participants au sondage en lien avec les sciences géographiques**

Nous observons une proportion égale des participants dans chacun des groupes. Ce sondage étant anonyme, nous ne nous pouvons donc pas faire un constat de l'expertise des contributeurs en regardant, par exemple, leur implication dans le projet sur les listes de diffusion OSM et leur nombre de contributions. De manière générale, nous essayons de définir un contributeur « expert » comme celui qui est abonné à au moins une des listes de diffusion OSM et y participe activement. Un contributeur « expert » OSM est aussi identifié par son implication dans le projet et par la connaissance générale qu'il possède du projet (ex : les tags OSM, les différents modes de contribution et les outils de correction et leurs adéquations selon les circonstances, le contenu du wiki OSM, etc.). Il connaît aussi en profondeur et est en charge de certaines activités spécifiques du projet (ex : la traduction en français de la documentation anglaise dans le wiki, la cartographie des nouveaux quartiers, la correction des erreurs faites par les nouveaux arrivants, la contribution d'informations sur une thématique spécifique comme les stations de ski ou les forêts, etc.).

### 2.2 Deuxième question : modes de contribution

Le but de cette question est de connaître les modes de contribution les plus utilisés sur OSM. Les réponses sont résumées dans le Tableau 25.

| | |
|---|---|
| Trace des images satellites | 56 (96,6 %) |
| Chargement de sources externes | 41 (70,7 %) |
| Collecte et téléchargement de traces GNSS | 39 (67,2 %) |
| Correction d'erreurs | 34 (58,6 %) |
| Renseignement des informations thématiques | 6 (10,3 %) |

**Tableau 25 : les modes de contribution les plus utilisés sur OSM**

Nous observons que le tracé des images satellites est le mode de contribution préféré des participants. Le chargement de sources externes, notamment les données issues du cadastre français et les données libres (*open data)*, est assez fréquent, de la même manière que la collecte et le téléchargement de traces GNSS. La correction d'erreurs est également un mode populaire.

D'autres modes de contribution ont été mentionnés :

*walking papers*,
relevés sur le terrain,
photos géolocalisées,
édition avec JOSM,
vérification d'infos par comparaison d'autres sources (photos, annuaires...),
travail avec les données obtenues auprès des collectivités,
tracé dans OSM depuis relevé GNSS mais sans *upload* du GPX,
repérage terrain par photo,
retranscription de mes observations réelles en ce qui concerne l'utilisation des terrains (landuse=*) ou des sites remarquables (fontaines, moulins),
conversion de données O*pen data* pour JOSM,

enquête terrain avec photo géolocalisées,
correction d'erreurs via Osmose,
analyse de donnée via SQL pour amélioration,
contribution au wiki,
collecte de photo sur le terrain,
collecte "manuelle" sur le terrain,
collecte de POI lors de promenades en ville (de mémoire, ou avec notes sur *walking papers*).

### 2.3 Troisième question : satisfaction à des outils d'édition

La troisième question du sondage est divisée en deux parties. La première partie porte sur la satisfaction globale des outils d'édition collaborative disponible sur OSM. Un résumé de ces résultats est présenté dans le Tableau 26.

| Satisfaits | Non-satisfaits | Pas d'avis sur le sujet | Réponse vide |
|---|---|---|---|
| 39 | 2 | 14 | 3 |

**Tableau 26 : satisfaction des participants des outils d'édition collaborative OSM**

Trente-neuf participants ont donné un avis positif sur au moins un des outils fournis par OSM en répondant par "oui", "satisfait", "aucun problème", "très utile", "très complet" , "assez simple", "très abouti" ou encore "particulièrement adapté". Deux participants ont exprimé la non-satisfaction des outils OSM en répondant par "non" ou "trop technique". Quatorze participants n'ont pas exprimé d'avis sur leur satisfaction vis à vis des outils OSM. Trois participants n'ont pas fourni de réponse (réponse vide). De plus, trois participants ont fait des remarques négatives sur le projet de cartographie collaborative Wikimapia, mentionné aussi dans cette question. Ils ont vivement critiqué la violation des licences des données sources, notamment sur la couche Google Maps utilisé pour tracer des géométries dans Wikimapia. La deuxième partie de la question porte sur les fonctionnalités manquantes et les difficultés des outils actuels d'édition. La Figure 94 illustre la synthèse des réponses sous la forme d'une carte conceptuelle.

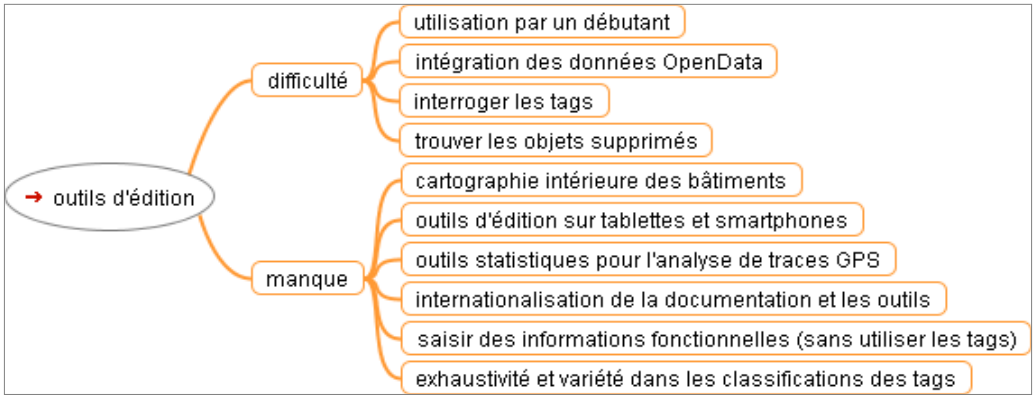



**Figure 94 : carte conceptuelle concernant les besoins des outils d'édition collaborative**

Les contributeurs ont exprimé des difficultés quant à l'utilisation des outils d'édition disponibles à l'heure actuelle. Plus particulièrement, ils ont témoigné la difficulté d'utilisation pour un débutant, notamment en ce qui concerne l'édition du contenu OSM. Par exemple, un participant indique : "OSM reste trop technique. Il faudrait pouvoir faire abstraction des éléments de bas niveau et pouvoir saisir directement des informations fonctionnelles sans utiliser les tags". D'autres difficultés sont plus spécifiques comme la difficulté d'intégrer l'open Data. En effet, un module JOSM permet la lecture de fichiers de données publiés en open data. Néanmoins, l'intégration dans le contenu OSM est entièrement manuelle. L'interrogation des tags afin de naviguer dans le système de tags OSM est signalée comme une tâche difficile. En effet, la documentation sur les tags OSM est entièrement en texte libre et il n'y a aucun mécanisme pour découvrir des tags potentiellement pertinents. Une dernière difficulté est de trouver des objets déjà supprimés dans la base. En réalité, ces objets ne sont jamais complètement effacés de la base de données OSM, mais ils ne sont pas pour autant accessibles. Un historique d'édition comme celui proposé dans cette thèse pourrait facilement permettre la consultation de ce type d'objets.

Parmi les fonctionnalités manquantes, les contributeurs signalent le besoin d'un outil pour cartographier l'intérieur des bâtiments, d'outils de saisie sur des *smartphones* ou tablettes, d'outils statistiques pour traiter les traces GNSS, l'internationalisation de la documentation, et des interfaces des outils d'édition.

Enfin, certains critiquent le système de *tagging* proposé actuellement par OSM, car ils voudraient pouvoir saisir des informations fonctionnelles sans devoir utiliser les tags, c'est-à-dire, un mécanisme qui permette de s'abstraire du modèle de tags. L'utilisation d'un vocabulaire formel est indiquée à ce propos. Les contributeurs voudraient aussi varier les thématiques proposées dans OSM, une méthode de découverte de vocabulaire comme celle proposée dans cette thèse pourrait aider à ce propos.

### 2.4 Quatrième question : renseignement de relations et propriétés

La quatrième question consistait à savoir si les contributeurs seraient prêts à renseigner des caractéristiques remarquables sur les objets, par exemple : "c'est le plus grand immeuble de la zone". De la même manière, il s'agissait de savoir s'ils seraient prêts à renseigner des relations "typées" entre certains objets du contenu OSM, de même que des relations entre des objets OSM et des objets d'un référentiel de données, par exemple "un restaurant d'OSM est au rez-de-chaussée d'un bâtiment du RGE®", "l'abribus d'OSM est exactement en face du bureau de poste du RGE®". Un résumé des avis exprimés sont présentés dans le Tableau 27.

| En accord | En désaccord | Assez d'accord | Question non comprise | Pas d'avis sur le sujet | Réponse vide |
|---|---|---|---|---|---|
| 7 | 27 | 9 | 4 | 3 | 8 |

**Tableau 27 : avis sur le renseignement de relations et propriétés des objets**

Sept participants sont en accord avec le fait de renseigner des informations sur les relations et propriétés des objets et vingt-sept participants ont exprimé leur désaccord (“non, absolument pas”, “semble limitée”, “informations trop volatiles dans le temps”, “pas grand intérêt”, “trop dépendant du contexte”). Neuf contributeurs sont quand même ouverts et donc assez en accord avec la proposition (“je ne doute pas que certains trouverons ça utile pour l'interrogation peut-être”, “pour pallier un manque temporaire d'information précise”, “pourquoi pas”, “ce type d'information peut faciliter l'orientation”, “ça doit probablement faciliter le repérage”, “ce qui pourrait rendre le contenu plus utile sont des contenus totalement subjectifs”). Trois participants n’ont pas fait de commentaires pertinents autour de ce sujet. Enfin, huit participants n’ont pas donné de réponses. La Figure 95 illustre la synthèse sur le contenu des réponses sous la forme d’une carte conceptuelle.

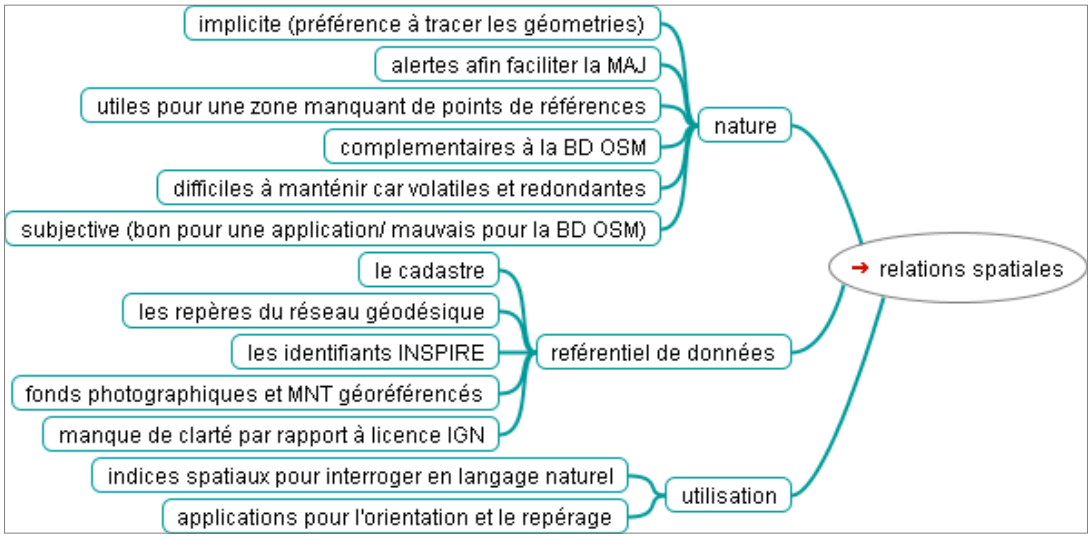



**Figure 95 : carte conceptuelle concernant le renseignement des relations spatiales**

Les contributeurs donnent d’abord leur interprétation sur ce qu’est une relation spatiale, puis plus précisément sur sa nature. Pour un contributeur, une relation spatiale est fréquemment implicite car elle ne peut être dérivée à partir des géométries. En effet, les contributeurs expriment la préférence de tracer des géométries. Une relation spatiale explicite est considérée utile pour des zones où il manque des points de référence. Elle peut servir d’alertes pour faciliter la mise à jour, un contributeur opine “cela facilite aussi la maintenance des données, dans le cas où un nouvel immeuble est construit ou que l'abribus est déplacé”. Néanmoins, de nombreux contributeurs ont signalé que les relations explicites devaient être conservées indépendamment de la base de données OSM. Dans le cas contraire, il est très difficile de les maintenir car ces informations peuvent être volatiles et devenir redondantes. A ce propos, un contributeur indique : “le fait que peu d'objets soient liés permet de gérer facilement les modifications observées. Si chaque objet était lié il y aurait des problèmes à chaque édition”. Certains contributeurs considèrent que ces informations sont très subjectives, un point négatif pour certains mais plutôt positif pour d’autres.

Par rapport à l'utilisation d'un référentiel de données IGN, les contributeurs utilisent actuellement le cadastre français comme source principale de référence. Ils signalent le besoin d'avoir accès aux repères du réseau géodésique, aux identifiant INSPIRE ainsi qu'à des fonds photographiques et MNT géoréférencés. Le manque de clarté dans la licence d'utilisation des données IGN les empêche de s'en servir plus.

Les contributeurs signalent que les relations spatiales explicites ont un grand intérêt à être utilisées comme des indexes spatiaux afin de permettre l'interrogation (et précisément en langage naturel) des données. A ce propos, un contributeur indique : "des indices spatiaux comme *c'est à côté de l'église, de la poste,* fonctionnent toujours très bien car une fois trouvé cet objet important (à l'aide d'une personne du coin par ex), on réduit son champ de recherche et/ou on attend la personne sur ce lieu de rendez vous".

Les relations spatiales montrent un intérêt pour les applications d'orientation et navigation, un contributeur indique "une application pour *smartphone* serait plus utile si elle pouvait indiquer les itinéraires en donnant des repères très visuels : *l'abribus se trouve juste en face du bureau de poste* est un excellent exemple".

Une semaine après l'envoi du sondage, un contributeur nous a demandé sur la liste de diffusion s'il pouvait en savoir plus sur la question des relations spatiales car il n'avait pas bien compris. Nous avons laissé les autres contributeurs répondre à cette question afin d'observer l'échange d'idées qui en a découlé. L'idée importante qui a surgi de cette échange est qu'il faudrait pouvoir renseigner des relations remarquables entre objets en définissant un nouveau tag ou de les dériver à partir des géométries existantes. Les contributeurs discutaient de cette idée dans le cadre d'une application de routage qui aurait besoin d'une base de relations en intégrant des points de repères locaux pour produire des descriptions pertinentes d'un itinéraire. Les types de relations discutées sont celles liées à la visibilité d'un objet du point de vue d'un autre objet, par exemple "un MacDonald qui est visible depuis l'entrée d'une autoroute" ou "deux restaurants côte à côte indiscernables par leurs tags usuels peuvent très bien avoir une différence importante dans leur apparence". D'autres types de relations discutées sont celles liées à l'isolement d'une entité, par exemple "la maison bleue au milieu des maisons blanches".

### 2.5 Cinquième question : conflits d'édition collaborative

La cinquième question consistait à identifier quels pouvaient être les conflits d'édition collaborative entre les contributeurs OSM. Les avis des contributeurs sur ce sujet sont présentés dans le Tableau 28.

| Oui | Non | Pas de réponse | Réponse hors sujet |
|---|---|---|---|
| 27 | 25 | 5 | 1 |

**Tableau 28 : conflit d'édition expérimenté par les participants**

Vingt-sept participants ont indiqué qu'ils avaient expérimenté des conflits d'éditions, vingt-cinq n'ont jamais expérimenté un conflit d'édition collaboratif et cinq n'ont pas répondu. Un participant n'a pas expérimenté un conflit d'édition mais un autre type de conflit (à savoir, du codage des objets). Plus précisément, la Figure 96 illustre la synthèse du contenu des réponses sous la forme d'une carte conceptuelle.

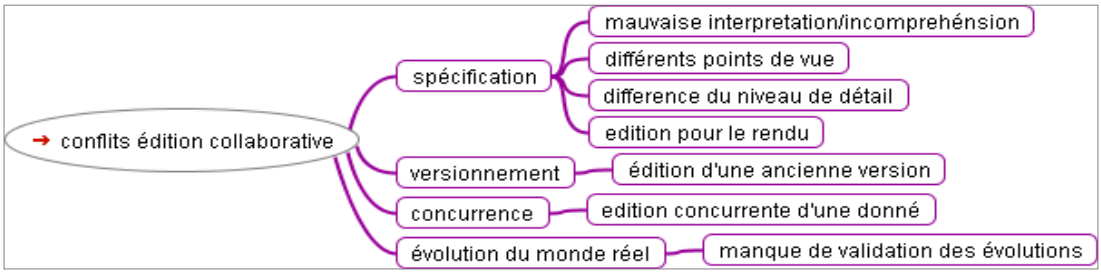



**Figure 96 : carte conceptuelle concernant les conflits d'édition collaborative**

En résumé, les conflits d'éditions collaboratives signalés peuvent être classifiés en quatre catégories principales : ceux liés à la **spécification**, au **versionnement**, ceux liés à l'**édition concurrente**, et enfin ceux liés à des **évolutions du monde réel**.

Dans le cas d'un conflit lié à la spécification, il existe quatre sous-catégories. Des exemples signalés par les contributeurs sont présentés ci dessous :

1) **Mauvaise interprétation / incompréhension** : "Très rapidement confronté au problème du tag des commerces. Doit-on créer un nœud pour définir le commerce, ou le commerce est-il directement lié au bâtiment ? Dans ce cas, comment tagger les bâtiments abritant commerce et habitation".
2) **Différents points de vue** : "Différence d'interprétation d'une zone humide : moi, qui habite à côté, ai mis l'étiquette "marécage" ; un contributeur lointain, au vu des photos satellite, y voyait une "prairie". Il tenait à sa "prairie" de manière crispée ; j'ai finalement eu gain de cause", et aussi "Le seul vrai conflit que j'ai connu était sur la caractérisation d'une voie que je considérais comme "*primary*" et qu'un autre contributeur modifié en "*secondary*" : nous avons trouvé un compromis sous la forme de 2 portions (une primaire, l'autre secondaire)".
3) **Différence dans le niveau de détail** : "Un doublon entre un point unique et un polygone, par exemple un point pour indiquer la présence d'un parking et le polygone qui définit son emprise".
4) **Edition pour le rendu** : "Un utilisateur (que je connais dans le monde réel) qui a supprimé des pistes cyclable sur un pont près de chez lui car le rendu n'était pas joli, comme s'il y avait plusieurs pont alors que c'est un pont commun avec la rue".

Des conflits de **versionnement** sont également indiqués comme celui-ci indiqué par un contributeur : "L'édition d'une version plus récente sur le serveur que chez moi en local". Des conflits d'**édition concurrente** peuvent se produire, à ce propos un contributeur explique : "il

arrive d'avoir un message indiquant que quelqu'un d'autre a modifié des objets en même temps". D'autres conflits correspondent au **non suivi des évolutions du monde réel**, comme les conflits qui arrivent toujours entre un contributeur qui travaille « en chambre » (depuis des données cadastrale, ou traces des images satellites Bing ou autre) et un contributeur sur le terrain".

### 2.6 Conclusion générale

Nous retenons plusieurs points importants de ce sondage :

- 55% de contributeurs ont une formation en lien avec les sciences géographiques. Il faudrait interroger les 45% de contributeurs restants afin de connaître leurs motivations à contribuer,
- les contributeurs n'ont pas de limites dans les modes de contribution à OSM. Ils préfèrent dessiner à partir de données issues d'images satellites. De fait, il serait sans doute intéressant d'améliorer les outils disponibles liés à ce mode de contribution. Il apparaît également important de remarquer qu'une bonne partie des contributeurs (58%) participent à la correction d'erreurs qui est un travail nécessitant d'être rigoureux,
- les contributeurs OSM sont assez satisfaits des outils de saisie existants proposés par la communauté OSM. Néanmoins, ils signalent des difficultés et des fonctionnalités manquantes dans ces outils de saisie, en particulier au niveau des nouveaux contributeurs. Ceux qui ont une bonne connaissance des outils expriment le besoin d'améliorer certains aspects comme par exemple celui de faciliter l'intégration de l'open data et d'améliorer le système de tags OSM (pas assez exhaustif et difficile à interroger),
- les contributeurs expriment la préférence de tracer des géométries sur la carte au lieu d'expliciter des relations spatiales. Ils indiquent néanmoins que cela peut faciliter la mise à jour car elles peuvent servir d'alertes. Elles peuvent aussi être utiles pour des zones où il manque des points de référence. De manière unanime, ils signalent que les relations spatiales montrent un intérêt pour les applications d'orientation et de navigation basées sur des données OSM. Dans tous les cas, elles doivent rester indépendantes de la base de données OSM,
- plus de la moitié des contributeurs ont déjà expérimenté un conflit d'édition collaborative. Les conflits sont en majeure partie liés à des problèmes de spécifications. Les contributeurs indiquent moins de conflits liés à l'édition concurrente, au versionnement et aux évolutions. Néanmoins, plusieurs d'entre eux signalent que ces sont ici des problématiques importantes et qui le deviendront de plus en plus à court terme.

Ce sondage nous a permis de connaître les points de vue des contributeurs OSM France sur différents sujets : leurs besoins d'outils de saisie, leur perception à propos de l'utilité des relations spatiales, certaines de leurs attentes vis-à-vis de l'IGN et enfin les conflits d'édition les plus fréquents qu'ils ont rencontrés. Certains aspects comme les « spécifications » OSM, l'édition concurrente et l'utilité des relations spatiales sont considérés dans ma proposition de thèse et sont développés en détail dans le manuscrit.

# Annexe B – Réponses au sondage proposé à des contributeurs OSM

Cette annexe présente les réponses précises aux trois dernières questions (#3, #4, #5) du sondage proposé aux contributeurs d'OSM.

**Réponses à la question "Êtes-vous satisfait de la façon dont vous pouvez saisir des données dans OpenStreetMap ou des outils d'édition collaborative similaires (ex : Wikimapia) ? Si vous pensez à quelque chose qui vous semble particulièrement utile ou quelque chose qui manque et serait apprécié, pouvez-vous le mentionner ?"**

**Legende pour la classification des réponses :** avis positif, avis négatif, pas de réponse,un manque/une amélioration, reference negative à wikimapia.

**Sondage#1 :** Gestion des "layers" pour les bâtiments à plusieurs étages. J'ai entendu parler d'un projet de "Indoorsmap" (Université de Heidelberg, je crois), qui répondrait à cette carence. Possibilité d'éditer les données OSM de type tel, adresse, à partir d'autres interfaces publiques de type Boussole, avec authentification via OSM OAuth.

**Sondage#2 :** Oui

**Sondage#3 :** Oui, et les outils etant libre il évolue aufildes jours et des besoins

**Sondage#4 :** Je suis satisfait des outils car ils sont complets, et évoluent rapidement pour améliorer leur clarté, proposer des outils plus puissants. Toutefois j'admets qu'il faut "s'accrocher" pour les débuts (installation, configuration ...). J'ai directement aidé un ami à créer son compte et installer / configurer JOSM. C'est désormais un contributeur très actif. Il m'a depuis indiqué que sans le coup de pouce initial, il n'aurait probablement pas contribué à OpenStreetMap

**Sondage#5 :** Nous disposons de plus en plus de données publiques (open data) mais nous n'avons pas les ressources pour les intégrer dans OSM, ni les outils pour les mises à jour. Un regret est aussi la communication en sens unique de ces données publiques qui contiennent souvent des erreurs ou des retards. Hors, il est impossible de signaler ces erreurs auprès de leurs auteurs, ce qui rend notre synchronisation encore plus compliquée. Quelque chose d'utile serait aussi de disposer des fonds photographiques (et MNT) de l'IGN gratuitement et déjà géoréférencés. Je suis sûr que l'IGN gagnerait à collaborer avec OSM plutôt que d'ignorer le public des amateurs passionnés de géographie. Le milieu de l'astronomie, par exemple, a depuis longtemps fait tomber ces barrières et préjugés entre professionnels et amateurs éclairés.

**Sondage#6 :** Le système de tag est suffisamment souple, ça me satisfait pleinement. La seule difficulté est de trouver le tags correspondant à l'objet observé mais le wiki, les presets JOSM et d'autres initiatives (telles qu'Osmecum) comblent le besoint.

**Sondage#7 :** Globalement satisfait mais il manque des outils de suivi personnalisés

**Sondage#8 :** Manque la possibilité de tracé des lignes et polylignes sur les tablettes à écrans capacitifs (du aux limites technologiques). Manque aussi une application pour smartphone permettant de tester la routabilité de de l'endroit où on navigue, incluant un système de correction. NavFree peut permettre de le faire, mais ce n'est pas pratique, parce que pas prévu pour cela.

**Sondage#9 :** oui

**Sondage#10 :** oui

**Sondage#11 :** OSM reste "trop" technique. Il faudrait pouvoir faire abstraction des éléments bas niveau et pouvoir saisir directement des informations fonctionnelles sans utiliser les tags.

**Sondage#12 :** Globalement oui. Il manque dans JOSM la possibilité de reprojection à la volée. Pour le moment, il faut adapter la projection à la couche source (planche cadastrale, imagerie Bing, etc.) Du coup, on ne peut pas superposer des données cadastrales avec une orthophoto et digitaliser de nouveaux objets en confrontant plusieurs sources.

**Sondage#13 :** Pas de réponse

**Sondage#14 :** Les outils à dispositions sont excellents, je préfère tout de même JOSM qui fonctionne sur toutes les plateformes. J'ai créé un petit tutoriel pour inciter les débutants à utiliser JOSM, vous le trouverez à l'adresse http://www.partir-en-vtt.com /php/articles/voir_article.php?id_article=282

**Sondage#15 :** oui

**Sondage#16 :** oui

**Sondage#17 :** Je suis totalement satisfait par la façon dont je peux saisir les données. Ce qui m'intéresse le plus c'est la fait de trouver un activité informatique formatrice pour mon enfant. Il est devenu accroc à OSM et commence sérieusement à faire de la saisie depuis un an. La première modification de l'une de ses contributions par un autre contributeur était un événement.
Il en était heureux comme si il y avait un reconnaissance de son travail pour ses paires et à aucun moment il a râlé, il a même était vérifier sur le terrain que la modif était justifiée. :-)

**Sondage#18 :** Une chose qui manque est la possibilité de reprojeter des images raster à la volée pour les superposer quand elles n'ont pas la même projection. L'exemple typique est la planche cadastrale en Lambert zone ou le WMS du cadastre vectorisé en conique conforme 9 zones, qu'il serait pratique de superposer à une orthophoto en pseudo-Mercator "EPSG:900913"
du type Bing, ou Orthophoto du Craig (sur l'Auvergne) par exemple.

**Sondage#19 :** Globalement satisfait. Après, il manque toujours un petit quelque chose sur un problème ponctuel. Et la solution peut exister sans qu'on le sache (plugin méconnu, autre logiciel d'édition...) J'aimerais bien un outil pour faire des moyennes de trace GPS à partir du stock disponible dans OSM. Pour retracer plus précisément les grands axes routiers.

**Sondage#20 :** Satisfait d'OSM. Pas du tout du comportement de wikimapia (problème de licence). La cible du questionnaire sont les utilisateur expérimenté, donc en tant que tel j'ai tendance à mettre en place moi même ce qui me manque, sauf si ça prend vraiment trop de temps.

**Sondage#21 :** Satisfait par le projet OpenStreetMap.

**Sondage#22 :** * Francisation de Potlatch * Mise en avant des outils de contrôle qualité * Traduction de la doc

**Sondage#23 :** Plutôt satisfait des deux principaux éditeurs d'OSM : JOSM et Merkaartor. La partie la plus délicate dans l'interface utilisateur est sans doute la gestion des conflits d'édition. Mais c'est un point intrinsèquement difficile.

**Sondage#24 :** Très satisfait par JOSM. Il manque des outils plus adaptés à la saisie sur le terrain (smartphone, tablette), surtout pour de la correction d'erreur.

**Sondage#25 :** Les outils de saisie OSM sont particulièrement adaptés malgré l'équipe de développement éclatée et réduite.

**Sondage#26 :** oui pour openstreetmap un wiki performant et bien hierarchisé serait mieux

**Sondage#27 :** Pouvoir retrouver facilement un objet qu'on a créé et qui a manifestement été supprimé, de manière à pouvoir identifier et échanger avec le contributeur auteur de la suppression.

**Sondage#28 :** oui

**Sondage#29 :** Oui, Je trouve le concept très facile à prendre en main.

**Sondage#30 :** Oui, mais pourrait être mieux au niveau de la gestion des relations (au sens OSM), c'est à dire un modèle de donnée faisant référence à d'autres éléments pour former des éléments logique de plus grande taille

**Sondage#31 :** J'utilise JOSM, que je considère comme un outil très abouti ; l'outil d'aide à l'import des données du cadastre vectorisé (qui y est incorporé sous forme de "plugin") évolue très fréquement pour s'améliorer ; il est déjà productif, mais j'attends avec impatience les améliorations en cours

**Sondage#32 :** Le mode de contribution pour OSM est assez simple, mais manque de concertation au niveau des tags. Wikimapia utilise des tuiles de Google Maps pour permettre de cartographier, et ce, sans l'accord de Google Maps.

**Sondage#33 :** Satisfait

**Sondage#34 :** Il manque des outils vraiment simples et utilisables par des débutants pour dessiner un plan de quartier ou un plan d'intérieur (de centre commercial, par exemple). Par exemple, on voudrait demander à des salariés de réaliser un plan d'intérieur de leurs bureaux, pour faciliter l'accès par des visiteurs handicapés, ou bien le plan d'intérieur d'une station de métro. Mais Potlatch ne permet pas de le faire facilement et JOSM est trop compliqué à prendre en main car il oblige à connaître les conventions de tagging. Il faudrait des logiciels Web de saisie d'une simplicité similaire aux logiciels grands publics destinés à dessiner des plans de cuisine pour faire des aménagements intérieurs.

**Sondage#35 :** Oui

**Sondage#36 :** limitation de download des api

**Sondage#37 :** Je n'utilise pas Potlatch, car trop basique à mon goût. Seul réel défaut est qu'il arrive (ou arrivait) souvent que les nouveaux n'utilisent pas correctement l'outil pour tracer des lignes (= routes, chemins). Ceci a pour conséquence d'avoir des routes non connectées dans OSM, alors qu'elles s'affichent correctement dans Potlatch. Il existe heureusement plusieurs outils de contrôle de qualité pour trouver et corriger ce genre d'erreurs. JOSM est un outil très complet, peut-être pas à la portée de tous, mais indispensable pour toute personne voulant s'investir un peu plus dans le projet que le simple ajout de points d'intérêts ou d'un nom de rue.

**Sondage#38 :** Pouvoir enrigistrer l'orientation (azimut & elevation) en plus des coordonnées GPS lors de la prise d'une photo sur un smartphone Android.

**Sondage#39 :** Oui pour OSM mais non pour wikimapia. JOSM est maintenant selon moi un véritable logiciel SIG dédié édition de données OSM.

**Sondage#40 :** Satisfait : Oui Amélioration : pleins Offrir la possibilité de varier les thématiques. OSM s'applique bien à certaines couches d'information (routes, bati, occupation du sol). Son

usage pourrait être étendu à d'autres thématiques qui impactent également la gestion du territoire : - zones humides : la liste des Tags est un peu courte pour cette préoccupation actuelle majeure - pédologie : la combinaison ponctuels/polgyones qui s'apparente aux relevés et à la restitution carto s'intégrerait bien dans un outil comme JOSM - relevés faunistiques et floristique ... Ces évolutions ne dépendent pas vraiment des outils mais plutôt d'un manque de détail dans les classifications proposées à l'heure actuelle dans OSM. Par contre leur acquisition serait facilitée par la disponibilité d'autres types de données dans les outils (imagerie avec d'autres spectres, Lidar, Relief et facteurs dérivés, ...)

**Sondage#41 :** oui.

**Sondage#42 :** Non. Il faudrait des éditeurs spécialisés (vélo, commerces etc.)

**Sondage#43 :** La solution OSM est très utile mais pour l'instant très complexe à expliquer à un néophyte. Une simplification des outils pourrait aider à rendre l'utilisation de OSM plus simple même pour une personne qui a déjà fait pas mal d'éditions.

**Sondage#44 :** Pas de réponse

**Sondage#45 :** Globalement satisfait. J'utilise josm qui fonctionne plutôt pas mal, mais il me semble qu'il manque quelques outils pour faciliter la tâche d'édition (notamment lors du travail sur l'occupation des sols).

**Sondage#46 :** Globalement, oui. Il manque la possibilité de filtrer certains éléments soit au téléchargement, soir à l'édition (couche occupation du sol, réseau routier, couche bati...)

**Sondage#47 :** Je suis satisfait. Les mailinglist sont très efficaces pour de l'aide ou des évolutions, et pour tout échange en général.

**Sondage#48 :** Oui, la simplicité d'utilisation des outils rend plaisantes les contributions à OpenStreetMap.

**Sondage#49 :** Une API JavaScript, pour multiplier les applications de saisies de données spécifiques. Par exemple créer une application dédiée à la saisie de données de Transport en Commun, des applications dédiés à l'exploitation des données OpenData.

**Sondage#50 :** Oui, JOSM marche très bien, et Potlach2 dépanne. Serait utile d'avoir un bon lien entre la doc et JOSM pour tagguer plus efficacement (noms de tags)

**Sondage#51 :** JOSM est un peu trop complexe pour le débutant, Potlach est trop brouillon AMHA. POI Collector un peu trop limité, il manque des outils plus grands publics.

**Sondage#52 :** oui à mon niveau de contribution cela convient bien, par contre, lorsque j'essaye de convaincre des utilisateurs non-sigistes de contribuer, ils ont du mal à se plonger dans les

logiciels, potlach est simple mais parfois difficile de charger des zones déjà bien mappées... pour qu'ils puissent rentrer qqs tags...

**Sondage#53 :** Pas de réponse

**Sondage#54 :** Il y encore du travail sur Vespucci pour le rendre vraiment utile.

**Sondage#55 :** Oui je suis satisfaite.

**Sondage#56 :** Il manque d'éditeurs plus simple et robuste sur mobile.

**Sondage#57 :** Aucun problème

**Sondage#58 :** Le relevé sur mobile est encore très imparfait et mériterait d'être exploré. Je pense par exemple à l'application Android de wheelmap.org qui fournit un très bon exemple (même s'il est limité à un domaine).

La Figure 97 présente le nuage de mots[78] construit à partir des réponses à la question #3 du sondage.


**Figure 97 : nuage de mots sur les réponses à la question #3 du sondage**

[78] http://www.wordle.net/create

**Réponses à la question “Pensez-vous que la description de caractéristiques de certains objets comme « c’est le plus grand immeuble de la zone » ou de relations entre certains objets comme « l’abribus est exactement en face du bureau de poste » peut rendre le contenu encore plus utile ? Si non, pourquoi ? Si oui, avez-vous des exemples en tête, en particulier des relations entre des nouveaux objets et des objets topographiques de référence comme ceux présents sur les cartes de sources gouvernementales comme l'IGN (routes, bâtiments, etc.) ?”**

**Legende pour la classification des réponses :** avis positif, avis négatif, assez d’accord, pas de réponse, question non comprise, intêrets possibles de renseigner des relations et des propriétés, reponse inutile, lien IGN, critiques du renseignement de relations et des propriétés

**Sondage#1 :** Je ne vois pas à quoi pourraient servir ce type d'information dans OSM, puisque la position relative des objets est fournie par la position des points sur la carte, et les objets peuvent être décrit par des attributs individuels (hauteur, nombre d'étage). En fait je ne suis pas sûr de comprendre l'intérêt de la question.

**Sondage#2 :** Non, Une carte suffit pour localiser et caract/riser les objets.

**Sondage#3 :** Pas de réponse

**Sondage#4 :** Ce genre de description ne me dérange pas. Mais je n'en saisis pas l'intérêt pratique. Si l'abribus est exactement en face du bureau de poste, je n'ai qu'à entrer ces deux éléments pour que cela soit mis en évidence. Les cartes de sources gouvernementales comme l'IGN ne sont pas une source autorisée pour entrer des informations dans la base OpenStreetMap. Je ne les utilise donc en aucun cas pour mes contributions.

**Sondage#5 :** Une des règles d'OSM est de "décrire le monde". Chacun y trouve son centre d'intérêt particulier (cyclisme, rando, etc) et motive ses renseignements particuliers. Certains se focalisent sur la cartographie des lignes à haute-tension, d'autres les lignes de chemin de fer ou les anciennes voies romaines, les monuments historiques, les zones d’atterrissage pour le vol à voile ou parapente, etc, etc . Ils pensent que ce qu'ils renseignent sera utile à d'autres personnes partageant les mêmes centres d'intérêt. C'est le principe du travail bénévole qui servira à d'autres qui se combine à la pratique d'un hobby personnel. Ceux qui décrivent les abris bus ou les pharmacies pensent sans doute que cela sera utile à d'autres parce qu'ils y trouvent eux-même de l'intérêt pour eux et pour les autres.

**Sondage#6 :** Étrange question... Le "contenu" OSM est utile parce qu'il est brut. Chacun peut imaginer d'exploiter les données de diverses manières. Déterminer (et éventuellement mettre en valeur) des relations entre objets est une exploitation possible des données OSM. Je ne doute pas que certains trouverons ça utile. J'imagine par exemple qu'une application pour smartphone serait plus utile si elle pouvait indiquer les itinéraires en donnant des repères très visuels : "l'abribus se trouve juste en face du bureau de poste" est un excellent exemple. Si la question est de faire rentrer explicitement ces relations dans la base de données, je pense que

c'est une très mauvaise idée. La base de données est plus saine avec des données brutes, tout ce qui peut être calculé devrait rester hors de la base.

**Sondage#7 :** Je pense que ce genre de relation est trop dépendante du contexte pour être rentrée dans la base de donnée OSM. En revanche un outil qui fournirait ce genre de description a partir des données OSM pourrait être très utile.

**Sondage#8 :** Le texte ne relève pas de la saisie primaire dans OSM, mais des applications dérivées qui sont ou seront capables de déterminer les positions relatives et de les commenter en langage naturel. Objet topographique de référence? En France, nous pouvons nous fier aux repères du réseau géodésique pour recaler la couverture de Bing. et les autres objets sont, en principe, calés grâce à cela.

**Sondage#9 :** non, car complètement inexploitable par des logiciels et assez abject. OSM n'est pas un roman, mais des objets avec une localisation géographique univoque. On ne va pas faire une base de données avec des positions relatives par rapport à d'autres objets, dont la position est encore relative à quelque chose, dont la position n'est peut être pas connue.

**Sondage#10 :** Non. C'est aux applications de prendre cet aspect en charge, en fonction des données de la base.

**Sondage#11 :** pour la requétage peut-être, pas pour la définition. Attention OSM est un modèle : la route n'est qu'un trait mathématique sans largeur matérialisée ; seuls les éléments définis par leur surface ont une "épaisseur" (mais pas de hauteur). Pour moi cela relève d'une possibilité d'interrogation en langage naturel : rechercher un élément proche de... ou en passant par tel endroit.

**Sondage#12 :** Non, ce ne sont pas des attributs objectifs d'un objet mais des attributs subjectifs du point de vue de l'utilisateur ou de l'application. Ces informations utiles au demeurant peuvent être dérivées par requêtage spatial et/ou attributaire. La description de telles caractéristiques d'un objet pourrait être vue comme une sorte de cache de requêtes spatiales antérieures. Il est du rôle d'applications tierces de définir l'ontologie de tels objets. Dans OpenStreetMap, les relations sont utilisées pour décrire, par exemple, des itinéraires de bus et ont leurs propres attributs (nom de la ligne, opérateur, référence, ...). On pourrait imaginer d'autres relations entre les différents bâtiments d'un lycée (cantine, salle de cours, administration, parking vélo, etc.)

**Sondage#13 :** Pas de réponse

**Sondage#14 :** Effectivement, des indices spatiaux comme c'est à côté de l'église, de la poste..fonctionnent toujours très bien car une fois trouvé cet objet important (à l'aide d'une personne du coin par ex), on réduit son champ de recherche et/ou on attend la personne sur ce lieu de rendez vous.

**Sondage#15 :** oui mais dans une seconde étape.

**Sondage#16 :** JNSP

**Sondage#17 :** je n'utilise pas ce genre d'objet

**Sondage#18 :** La taille des immeubles serait intéressante à renseigner de façon objective et la plus complète possible : indiquer quel est le plus grand immeuble n'a pas grand intérêt... pour quel usage ? Pour le positionnement relatif (abribus en face de la poste) c'est à la charge des outils externes d'analyse de la base de le calculer, ça ne doit pas être intégré dans la base "en dur". Un exemple concret : dans un logiciel de calcul d'itinéraire, on arrive à un rond point. Pour l'instant à l'entrée du rond-point un logiciel simple dira "prenez à droite, puis prenez la 2e à droite Rue machin". Un logiciel plus élaboré devra dire "au rond point, prenez tout droit Rue machin". C'est juste de l'analyse de données (les rond-points sont indiqués en tant que tels dans la base OSM) En ce qui concerne les relations entre les objets OSM et ceux des bases "officielles" comme celles d'INSPIRE, on attend que l'IGN fasse le premier pas et permette d'accéder à l'identifiant INSPIRE unique de chaque objet ! Quand cela existera, avec un moyen d'accès et de consultation aussi simple que l'url http://www.openstreetmap.org/browse/way/140456965 eh bien je crois que les contributeurs à OSM vont s'empresser de renseigner le lien dans les objets OSM ! Il n'y a qu'à voir les références externes qui existent déjà : *http://taginfo.openstreetmap.fr/keys/?key=ref%3AINSEE#map avec le code INSEE des communes *http://taginfo.openstreetmap.fr/keys/?key=ref%3Amhs#map avec le code indiqué dans la base Mérimée pour les Monuments Historiques * http://taginfo.openstreetmap.fr/keys/?key=ref%3Asandre#map avec le code du Sandre pour les cours d'eau On a même : * http://taginfo.openstreetmap.fr/keys/?key=ref%3ACEF#map avec le code utilisé par l'église catholique en France pour les lieux de culte ! Alors un ref:IGN ou ref:INSPIRE a des chances de se mettre en place comme une traînée de poudre. Pour moi la balle est dans le camp de l'IGN, avec un accès aux ressources (par pitié !) le plus simple possible (une url par exemple !).

**Sondage#19 :** Il est toujours difficile d'imaginer toutes les applications possibles alors ce genre d'info peut toujours être utile. Dans le cadre d'un guidage routier, plutôt que "tourner à droite à 100 m", "tourner à droite au prochain faux" ou "tourner à droite après le grand immeuble" sont des infos plus pratiques.

**Sondage#20 :** Pas de réponse

**Sondage#21 :** A la 1ère question ces caractéristiques me semblent être des informations trop "volatiles" dans le temps et donc difficiles à maintenir à jour. Je n'ai pas compris le sens de la 2ème question.

**Sondage#22 :** Non Si les données décrivant les caractéristiques propre d'un objet, et de ceux de sont entourage sont complètes, on doit pouvoir en déduire automatiquement ce type de caractéristiques.

**Sondage#23 :** Je suppose que les descriptions suggérées sont là pour pallier un manque temporaire d'information précise. Je ne pense pas que la donnée "approximative" soit beaucoup plus simple à indiquer que la donnée exacte.

**Sondage#24 :** Ce type d'information peut en effet être utile sur une zone manquant de points de référence. Si les objets de référence sont là (dans le cas d'OSM ce sont généralement les bâtiments issus du cadastre), il est facile et naturel de positionner les nouveaux objets pas rapport à ces objets de référence.

**Sondage#25 :** Utile pour l'utilisateur final qui cherche sa route. En dehors de cette utilisation particulière, l'intérêt de cette donnée me semble limitée. D'autre part, c'est une donnée qui peut être calculée par l'application cliente de la base de données.

**Sondage#26 :** non, car une carte bien faite sert à faire visuellement ou apres retraitement informtique toutes les relations necessaires, voulues par tous utilisateurs differents avec des centres 'intéret different.

**Sondage#27 :** Pourquoi pas mais cela devra être exploitable par des logiciels et services dérivés.

**Sondage#28 :** oui pour les relations

**Sondage#29 :** Non, Je pense que ce n'est pas encore le moment pour ce genre de tags. Il reste encore beaucoup trop à faire avant d'en arriver là. De plus, il sera plus difficile de modifier un objet simple s'il est en relation avec d'autres objets.

**Sondage#30 :** non, pas très utile. Si les coordonnées sont bien renseignées, on peut déterminer ça logiciellement.

**Sondage#31 :** non ; ce type d'information est directement fourni par le sig dès lors qu'il dispose des coordonnées de ces objets.

**Sondage#32 :** La description des immeubles avec des tags simple comme la hauteur, le type de toit, etc, peut permettre une représentation 3D fort sympathique. Mais est-ce très utile ? Pour exploiter le jeu de données dans des simulations, peut-être.

**Sondage#33 :** Non, j'ai du mal à en voir l'utilité. L'idée est à creuser, mais savoir qu'un bâtiment est en face d'un autre dans une base de donnée géographique, on ne devrait pas avoir besoin d'un tag ! C'est plutot un job de présentation des données, non?

**Sondage#34 :** J'aimerais pouvoir déterminer "quels sont les 2 ou 3 principaux repères géographiques à proximité d'un lieu donné". Ceci est nécessaire pour produire des plans d'accès à des bâtiments publics. En effet, il faut d'abord localiser le bâtiment cible puis, pour

choisir un périmètre de plan qui soit utile aux visiteurs, il faut choisir un périmètre qui soit suffisamment grand pour inclure un repère géographique très connu à proximité. Par exemple, pour faire le plan d'accès à la mairie de Boulogne-Billancourt, il est souhaitable d'y inclure le MacDonald's le plus proche et la station de métro la plus proche. Le problème, c'est que la notion de "repère principale" est subjective et est généralement connue des habitants. En leur demandant : pourriez-vous m'expliquer  comment me rendre dans ce lieu, ils utilisent des repères qu'ils estiment connus ou facilement repérables (Mac Do, métro, etc.).

**Sondage#35 :** Je ne sais pas

**Sondage#36 :** non, la base est géospatiale. - on peut trouver cette information grace aux données - cette information peut devenir inexacte

**Sondage#37 :** Absolument pas. Ce genre d'information pourrait être extrait de façon automatique, car si l'abribus est en face du bureau de poste, il sera affiché ainsi dans OSM. Quant au plus grand immeuble, ça se voit que c'est le plus grand (ou un des plus grands). Mais là encore, en calculant la zone que couvre le bâtiment en question et grâce à un tag comme building=yes, on pourrait déterminer de façon automatique quel est le plus grand bâtiment d'une zone prédéfinie.

**Sondage#38 :** Pas de réponse

**Sondage#39 :** A priori, non. Je décris simplement l'existant qui évolue peu.

**Sondage#40 :** Dans certains domaines, comme la randonnée, ce type d'information peut faciliter l'orientation. Si ces informations d'ordre topologiques font l'objet d'une typologie, pourquoi pas. Sinon, ca représente des remarques intéressantes à la lecture d'une carte, mais ca n'apporte aucune information structurée et exploitable pour améliorer/tester la qualité des données.

**Sondage#41 :** Pas de réponse

**Sondage#42 :** Non, ce genre d'information doit s'obtenir par analyse de la base de données : si l'immeuble est le plus grand, ça se mesure. Si l'abri bus est en face du bureau de poste, ça se constate.

**Sondage#43 :** Non. Il ne faut pas mélanger les informations sur le positionnement et les relations entre objets. En plus, actuellement, le fait que peu d'objets soient liés permet de gérer facilement les modifications observées. Si chaque objet était lié il y aurait des problèmes à chaque édition. L'ajout de la 3D dans OSM n'étant pour l'instant pas vraiment un objectif je vois mal comment gérer des relations comme l'objet est le plus grand. L'ajout pertinent d'informations tout autour est pour l'instant un bien meilleur objectif et prend déjà beaucoup de ressources.

**Sondage#44 :** Non. L'indexation d'autres sources mieux organisées peut lever des ambiguités de meilleure façon (notamment les bases de photos géolocalisées comme FlickR, et Wikimedia Commons, ou les URL vers des articles descriptifs avec des liens de référence utiles comme Wikipedia), avec l'aide aussi d'autres moteurs de recherche.

**Sondage#45 :** Je ne suis pas sur de comprendre la question. Les deux exemples d'attributs sont déductible de la base de données, il n'est pas nécessaire d'ajouter des informations redondantes.

**Sondage#46 :** Non. Si les données sont suffisamment complètes et précises, ce sont des indications qui peuvent être déduites : - si la hauteur des bâtiments est connue, il possible de déduire lequel est le plus haut - si l'abribus et la poste sont présents, il peut être déduit que l'abribus est en face de la poste Cela facilite aussi la maintenance des données, dans le cas où un nouvel immeuble plus est construit ou que l'abribus est déplacé.

**Sondage#47 :** ça doit probablement faciliter le repérage.

**Sondage#48 :** Non, les interprétations de données telles que la proximité ou la hauteur par rapport à d'autres bâtiments peut se faire via des applications tiers. De plus, ces données peuvent devenir invalides rapidement (déménagement de la Poste, construction d'un building) et donner lieux à des incohérences.

**Sondage#49 :** Non, ce genre d'information est importante pour orienté les gens

**Sondage#50 :** Oui, mais ces données sont difficiles à laisser interpréter par une machine. Trouver un moyen de fusionner ces informatins avec les données elle-mêmes pourrait être intéressant.

**Sondage#51 :** Pas de réponse

**Sondage#52 :** Oui cela serait sans doute utile, mais je pense qu'il y a bien suffisament de trous et de manques pour se consacrer aux objets avant de les décrire les uns par rapport aux autres

**Sondage#53 :** Pas de réponse

**Sondage#54 :** Pas de réponse

**Sondage#55 :** Oui c'est utile car ça ajoute une information.

**Sondage#56 :** Normalement non, les descriptions doivent être déductible grâce aux données de la base.

**Sondage#57 :** Non, car à partir du moment ou tout est correctement renseigné (position, hauteur des bâtiments...) il n'y a qu'a regarder les données.

**Sondage#58 :** Je ne suis pas sûr que les exemples cités soient pertinent car OSM étant une base, on devrait pouvoir déduire ces informations par du calcul. Ce qui pourrait rendre le contenu plus utile sont des contenus totalement subjectifs comme : "tour stalienne", "super restau", "endroit calme", "charmant", etc. (je suis conscient que ce n'est plus du tout dans le scope d'OSM).

La Figure 98 présente le nuage de mots construit à partir des réponses à la question #4 du sondage.


**Figure 98 : nuage de mots sur les réponses à la question #4 du sondage**

**Réponses à la question “Avez-vous déjà connu un conflit d’édition entre contributeurs ? Si oui, pouvez-vous le décrire ?”**

**Legende pour la classification des réponses :** avis positif, avis négatif, pas de réponse,conflits d’édition, autre type de conflit

**Sondage#1 :** Non, les quelques contacts que j'ai eu avec d'autres contributeurs concernant des erreurs ou désaccords dans mes contributions se sont toujours poursuivis très cordialement, avec atteinte d'un consensus.

**Sondage#2 :** Non

**Sondage#3 :** Oui, un désaccord sur des rage pour des voies semi piétonne en centre ville, mais après. Un échangea de mail nous avons trouvé un accord

**Sondage#4 :** Non. Le service de messagerie interne d'OSM est tout indiqué pour mettre en évidence des erreurs. On a relevé des erreurs que j'ai commises, et j'ai fait de même avec d'autres contributeurs. Dans la plupart des cas, à condition d'avoir une justification (page wiki, référence sur le terrain) les choses sont rapidement corrigées. Je n'ai pas encore vu la situation se pourrir. J'ai lu les échos d'une querelle concernant LE noeud de Jerusalem. L'enjeu était la langue utilisée pour l'attribut "name", donc le nom principal qui va s'afficher sur le rendu du site openstreetmap.org. Une querelle qui n'est finalement que pour un élément de façade : il est possible de faire un rendu en hébreu, un autre en arabe.

**Sondage#5 :** non

**Sondage#6 :** Ça m'est arrivé, oui. Difficile à décrire de manière générale : - J'avais renseigné des informations en les attachant au way d'un bâtiment, un autre contributeur a saisi les mêmes informations sur un node au dessus du bâtiment. C'est un cas fréquent. - J'avais saisi un sentier de randonnée. Un contributeur peu averti a télécharger une trace GPS, l'a converti en chemin et a valider sans plus de vérification. - Il y a eu d'autres cas, j'ai du mal à me souvenir des détails.

**Sondage#7 :** non

**Sondage#8 :** Il arrive d'avoir un message indiquant que quelqu'un d'autre a modifié des objets en même temps. Dans JOSM, il y a un module de résolution des conflits qui permet de fusionner les versions. Mais, je n'ai jamais été voir de quels objets, il s'agissait. Un conflit sur les attributs de l'objet? C'est parfois mis en discussion sur la liste de diffusion talk-fr et le peu dont je me souvienne concerne les chemins urbains (path ou footway?) Voir les archives de la liste en 2009 et 2010. Bon courage : 500 à 1200 messages/mois.

**Sondage#9 :** oui, une version plus récente sur le serveur que chez moi en local. Quoi dire de plus...?

**Sondage#10 :** Non

**Sondage#11 :** plusieurs cas différents : un objet défini initialement sous la forme d'un point que je peux transformer en une surface (apport ultérieur par le cadastre par ex.) ; ce n'est pas vraiment un conflit, mais une amélioration. Certains contributeurs ajoutent des éléments et/ou des formulations particulières pour que l'outil final donne un rendu particulier ; dans ce cas ils ne respectent pas nécessairement les standards du "modèle" général. Le seul vrai conflit que j'ai connu était sur la caractérisation d'une voie que je considérait comme "primary" et qu'un autre contributeur modifié en "secondary" : nous avons trouvé un compromis sous la forme de 2 portions (une primaire, l'autre secondaire) http://osm.org/go/0BNq74ZyW-- Un autre était un contributeur ignorant des modifications récentes (travaux de modification d'une voie d'accès) qui a annulé ma contribution abusivement. Après discussion, j'ai pu rétablir ma correction : http://osm.org/go/0BM~vnjau--?m

**Sondage#12 :** Je n'ai jamais été confronté à une telle situation. Je suis parfois en conflit d'édition avec un autre contributeur soit pour une utilisation de tags inapropriés (à mes yeux) soit parce que nous travaillons à plusieurs sur le même objet. OpenStreetMap permet de représenter une même réalité sous 2 formes différentes suivant le niveau de détail souhaité ou possible : point ou polygone. Une mairie, une école, un château d'eau peuvent exister sous les 2 forme. Si le bâtiment habritant la mairie, l'école ou le château d'eau, le point est remplacé par le polygone : il y précision et non pas conflit.

**Sondage#13 :** Pas de réponse

**Sondage#14 :** J'ai déjà eu un souci sur un chemin mais après discussion avec le contributeur, tout s'est bien arrangé. Je ne suis pas trop confronté à ce type de souci car je map dans des secteurs assez peu/pas couverts en données.

**Sondage#15 :** non

**Sondage#16 :** non

**Sondage#17 :** non

**Sondage#18 :** Oui, et cela s'est toujours résolu par une suppression d'autorité de la version la moins précise. Dans mon cas c'était toujours dû à une erreur d'inattention et pas de la mauvaise volonté (c'est le sens faible du mot conflit).

**Sondage#19 :** Plutôt des conflits sur le codage des objets. Exemple : sur des voies ferrées à espacement métrique, à tout repassé en écartement standard et n'a pas corrigé quand je lui ai signalé.

**Sondage#20 :** Cela peut même arriver pour un même contributeur. Ce n'est pas si grave en soit, 1- l'objet est décrit, c'est déjà un bon point. 2- Les outils de qualité arrivent, ou arriveront à

la détecter. Plus géant sont les conflits sur un même objet ou des contributeurs ne sont pas d'accord entre eux. Soit il n'ya pas de guerre et le dernier qui passe à raison (on suppose qu'il a conscience de qu'il fait, et le fait de bonne fois). OpenStreetMap est basé sur la bonne fois de contributeurs. Soit il y a une guerre et des modification en boucle et c'est plus gênant.

**Sondage#21 :** Non.

**Sondage#22 :** Non

**Sondage#23 :** De nombreuses fois : ça se passe soit en discussion, soit parce que le modèle le plus approprié finit par l'emporter. Un exemple : tracé des fossés à partir du cadastre, en mode filaire ou en mode "riverbank". Le mode riverbank est un abus, provenant de l'import automatique du cadastre. Il ne se justifie pas, et donne une fausse impression de précision. J'efface le riverbank, et j'envoie (souvent) un message à l'auteur de l'import pour expliquer la raison. Personne n'a contesté ni rétabli le riverbank.

**Sondage#24 :** Oui, mais souvent c'est un doublon entre un point unique et un polygone, par exemple un point pour indiquer la présence d'un parking et le polygone qui définit son emprise. Dans OpenStreetMap, ceci est détecté automatiquement par nos outils de contrôle qualité (osmose) et donc corrigé assez rapidement.

**Sondage#25 :** Oui, ça arrive souvent lors de cartoparties, lorsque plusieurs personnes travaillent sur une même zone ou une zone adjacente. Lorsque 2 contributeurs travaillent en même temps sur des zones proches, des objets étendus (routes, polygones de type d'occupation de sols, ou polygones de découpage administratifs) peuvent faire partie de ces deux zones, et être modifiés. OSM et les éditeurs prévoient cette situation, et proposent des outils de résolution de conflits. Hors cartoparties, les conflits sont rares.

**Sondage#26 :** non, puisque la guerre n'a finalement pas eu lieu. le conflit n'etait que sous jascent et bien traité par les parties, en toute cordialité (inteligence...?)

**Sondage#27 :** Il m'arrive de contacter des contributeurs qui laissent de nombreux points dupliqués à l'issue d'imports ratés de bâtiments du cadastre. Cela s'est toujours bien passé parce que je propose de les aider en signalant que ce problème m'est déjà arrivé. Ce n'est pas véritablement un conflit d'édition. Récemment, j'ai contacté un contributeur qui avait placé des panneaux de signalisation faisant référence à une nomenclature allemande. Il a été d'accord pour les remplacer par la référence française que je lui indiquait dans le wiki d'OSM.

**Sondage#28 :** oui deux modifications de routes dans un nouveau quartier. Plus exactement, une route crée sans conflit puis 2 personnes qui modifient cette route.

**Sondage#29 :** Oui, Très rapidement confronté au problème du tag des commerces. Doit-on créer un node pour définir le commerce, ou le commerce est-il directement lié au batiment ? Dans ce cas, comment tagger les batiments abritant commerce et Habitation.

**Sondage#30 :** Oui, mais tant qu'on peut discuter, on résoud en général le problème par la parole, quitte à se retrouver sur place pour confirmer des faits plutôt que des suppositions

**Sondage#31 :** oui. différence d'interprétation d'une zone humide : moi, qui habite à côté, ai mis l'étiquette "marécage" ; un contributeur lointain, au vu des photos satellite, y voyait une "prairie". Il tenait à sa "prairie" de manière crispée ; j'ai finalement eu gain de cause, l'usage à OSM étant de laisser le dernier mot à celui qui peut observer le terrain "de visu"

**Sondage#32 :** Le plus gros conflit d'édition que j'ai rencontré: Modification des données pour modifier le rendu, c'est une plaie.

**Sondage#33 :** non

**Sondage#34 :** Oui. Non, je n'ai plus le temps. :)

**Sondage#35 :** non

**Sondage#36 :** non - la probabilité que cela arrive est faible - un controle de qui fait quoi ou avant édition

**Sondage#37 :** Personnellement, non. Constaté des erreurs et contacté l'auteur, oui, une fois, et ça s'est passé sans problème puisque l'auteur était nouveau, ne comprenait pas encore tout à fait Potlatch et à rapidement corrigé ses propres erreurs (à savoir, interconnecter les routes, cfr question 3) en se basant sur les outils de contrôle de qualité.

**Sondage#38 :** Oui. Coincidence fortuite, la première rue que j'avais ajoutée l'avait aussi été par un autre utilisateur plutôt intensif (beaucoup de contributions) à peu de temps d'intervalle. Il en résulté un doublon, jusqu'à ce que je m'en rende compte et en supprime une. Une autre fois, c'est un utilsateur (que je connais dans le monde réel) qui a supprimé des pistes cyclable sur un pont près de chez lui car le rendu n'était pas joli (= comme s'il y avait plusieurs pont alors que c'est un pont commun avec la rue).

**Sondage#39 :** Oui, sur des classifications de route en tertiary/secondary

**Sondage#40 :** certains affectent un Tag à un ponctuel. d'autres créeront une meilleur représentation de l'objet pour ce même Tag. Pas vraiment un conflit, plus une question de précision. Les conflits portent plus souvent sur les Tags que sur la représentation géométrique des objets.

**Sondage#41 :** Pas de réponse

**Sondage#42 :** Oui, toujours entre un contributeur qui travaille « en chambre » (depuis des données cadastrale, bing ou autre) et un contributeur sur le terrain

**Sondage#43 :** Non

**Sondage#44 :** Pas de réponse

**Sondage#45 :** Pas de réponse

**Sondage#46 :** Quelques désaccords sur la classification de routes (secondary/tertiary ou tertiary/unclassified pour OSM)

**Sondage#47 :** Oui, mais pas souvent. C'est assez "paniquant" sur l'instant, mais au final on arrive à résoudre le conflit (j'utilise JOSM).

**Sondage#48 :** Non

**Sondage#49 :** Oui et non. Il m'est arrivé de devoir revenir en arrière sur des tags mais cela était dû soit à une méconnaissance d'un nouveau contributeur soit un problème de langage : faut-il utiliser color ou colour ?

**Sondage#50 :** Non.

**Sondage#51 :** Non.

**Sondage#52 :** non, pas encore

**Sondage#53 :** Pas de réponse

**Sondage#54 :** non et si je le voyais, je le corrigerais simplement

**Sondage#55 :** Non

**Sondage#56 :** non

**Sondage#57 :** non

**Sondage#58 :** Le cas classique est le polygone vs point.

La Figure 99 montre le nuage de mots construit à partir des réponses à la question #5 du sondage.

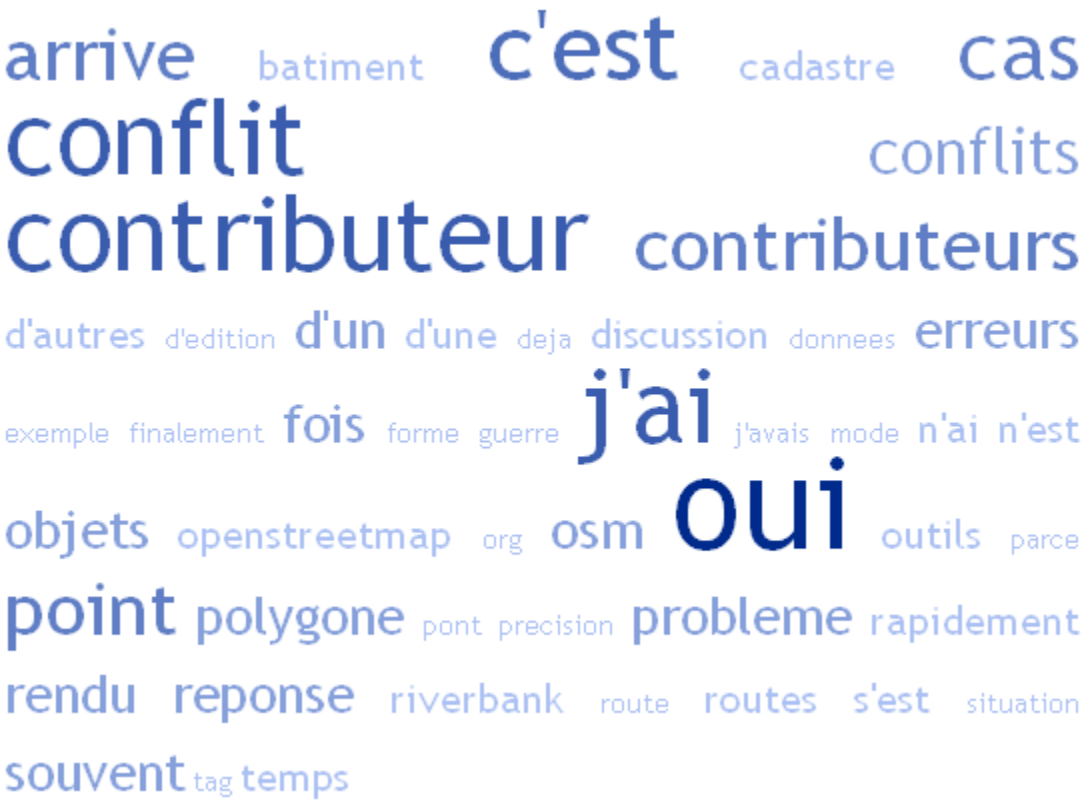



**Figure 99 : nuage de mots sur les réponses à la question #5 du sondage**

# Annexe C – Expérience avec OSMonto

Cette annexe liste les étiquettes des classes de notre version française de l'ontologie OSMOnto (Codescu et al. 2011) utilisée pour initialiser notre *catalogue d'éléments de vocabulaire formel*. Pour les 19 thèmes :

**k_amenity (Equipements publics)**

```
v_airport (aéroport@fr)
v_arts_centre (centre culturel@fr)
v_arts_centre (centre des arts@fr)
v_atm (distributeur automatique de billets@fr)
v_bank (banque@fr)
v_bar (bistrot@fr)
v_bench (banc public@fr)
v_bicycle_parking (parking à vélos@fr)
v_bicycle_rental (location de vélos@fr)
v_bureau_de_change (bureau de change@fr)
v_bus_station (gare routière@fr)
v_cafe (café@fr)
v_car_rental (location de voiture@fr)
v_car_sharing (station d'autopartage@fr)
v_car_wash (station de lavage pour automobiles@fr)
v_casino (casino@fr)
v_charging_station (station de recharge@fr)
v_cinema (cinéma@fr)
v_clinic (clinique@fr)
v_college (établissement d'enseignement supérieur non universitaire@fr)
v_courthouse (palais de justice@fr)
v_dentist (dentiste@fr)
v_doctors (cabinet médical@fr)
v_doctors (médecin généraliste@fr)
v_doctors (médecin spécialiste@fr)
v_drinking_water (source d'eau potable@fr)
v_embassy (embassade@fr)
v_emergency_phone (borne d'appel@fr)
v_ferry_terminal (terminal de ferry@fr)
v_fire_station (caserne de pompiers@fr)
v_food_court (zone de restauration@fr)
v_fountain (fontaine@fr)
v_fuel (station essence@fr)
v_grave_yard (petit cimetière@fr)
v_hospital (hôpital@fr)
v_kindergarten (école maternelle@fr)
v_kindergarten (jardin d'enfants@fr)
v_library (bibliothèque@fr)
v_marketplace (place de marché@fr)
v_nightclub (boîte de nuit@fr)
v_nursing_home (maison de retraite@fr)
v_parking (parking@fr)
v_parking (stationnement@fr)
v_pharmacy (pharmacie@fr)
v_place_of_worship (chapelle@fr)
```

v_place_of_worship (église@fr)
v_place_of_worship (mosquée@fr)
v_place_of_worship (synagogue@fr)
v_place_of_worship (temple@fr)
v_police (gendarmerie@fr)
v_post_box (boîte aux lettres@fr)
v_post_office (bureau de poste@fr)
v_prison (prison@fr)
v_public_building (bâtiment public@fr
v_pub (pub@fr)
v_recycling (point de collecte pour le recyclage@fr)
v_restaurant (restaurant@fr)
v_sauna (sauna@fr)
v_school (collège@fr)
v_school (école primaire@fr)
v_school (lycée@fr)
v_shelter (abribus@fr)
v_shelter (refuge de montagne@fr)
v_taxi (station de taxis@fr)
v_telephone (téléphone public@fr)
v_theatre (théâtre@fr)
v_theatre (salle de spectable@fr)
v_toilets (toilettes@fr)
v_town_hall (mairie@fr)
v_university (campus universitaire@fr)
v_university (université@fr)
v_veterinary (vétérinaire@fr)
v_waste_basket (poubelle publique@fr)

### k_shop (Commerces)

v_bakery (boulangerie@fr)
v_beauty (institut de beauté@fr)
v_beauty (salon de beauté@fr)
v_beverages (vente de boissons à emporter@fr)
v_bicycle (magasin de vélos@fr)
v_books (librairie@fr)
v_butcher (boucherie@fr)
v_butcher (charcuterie@fr)
v_car_repair (garage automobile@fr)
v_car (concessionnaire automobile@fr)
v_chemist (droguerie@fr)
v_clothes (boutique de vêtements@fr)
v_computer (boutique d'informatique@fr)
v_department_store (grand magasin@fr)
v_doityourself (magasin de bricolage@fr)
v_dry_cleaning (magasin de pressing@fr)
v_electronics (magasin électroménager@fr)
v_electronics (magasin électronique@fr)
v_florist (fleuriste@fr)
v_furniture (décoration d'intérieur@fr)
v_furniture (mobilier@fr)
v_garden_centre (jardinerie@fr)
v_hairdresser (coiffure@fr)
v_hardware (quincaillerie@fr)
v_hardware (serrurerie@fr)
v_jewelry (bijouterie@fr)

v_kiosk (kiosque à journaux@fr)
v_kiosk (tabac@fr)
v_laundry (laverie@fr)
v_mall (centre commercial@fr)
v_massage (salon de massage@fr)
v_optician (opticien@fr)
v_organic (magasin bio@fr)
v_pet (magasin pour animaux de compagnie@fr)
v_seafood (magasin de vente de fruits de mer@fr)
v_seafood (magasin de vente de poisson@fr)
v_shoes (chaussure@fr)
v_shoes (magasin de chaussures@fr)
v_sports (magasin de sport@fr)
v_stationery (papeterie@fr)
v_supermarket (supermarché@fr)
v_toys (magasin de jouets@fr)
v_travel_agency (agence de voyage@fr)
v_video (location de vidéo ou dvd@fr)
v_video (vente de vidéo ou dvd@fr)

### k_natural (Formation végétale)

v_bay (baie@fr)
v_beach (plage@fr)
v_cave_entrance (entrée de grotte@fr)
v_cliff (falaise@fr)
v_coastline (littoral@fr)
v_fell (montagne@fr)
v_heath (lande@fr)
v_mud (terrain boueux@fr)
v_peak (sommet@fr)
v_sand (dune@fr)
v_scrub (friche@fr)
v_scrub (garrigue@fr)
v_scrub (maquis@fr)
v_tree (arbre isolé@fr)
v_tree (arbre remarquable@fr)
v_valley (vallée@fr)
v_volcano (volcan@fr)
v_water (étang@fr)
v_water (lac@fr)
v_wetland (zone humide@fr)
v_wood (bois@fr)

### k_leisure (Loisirs)

v_arena (arènes@fr)
v_garden (jardin@fr)
v_golf_course (terrain de golf@fr)
v_marina (port de plaisance@fr)
v_miniature_golf (golf miniature@fr)
v_nature_reserve (réserve naturelle@fr)
v_park (parc@fr)
v_pitch (terrain de sport@fr)
v_playground (zone de jeu enfant@fr)
v_recreation_ground (aire de jeux@fr)
v_slipway (cale@fr)
v_sports_centre (centre sportif@fr)

v_stadium (stade@fr)
v_swimming_pool (piscine@fr)
v_track (terrain de course@fr)
v_water_park (parc aquatique@fr)

### k_tourism (Tourisme)

v_alpine_hut (refuge de montagne@fr)
v_artwork (oeuvre d'art@fr)
v_attraction (attraction touristique@fr)
v_camp_site (camping@fr)
v_caravan_site (aire pour caravanes@fr)
v_chalet (chalet@fr)
v_guest_house (chambres d'hôtes@fr)
v_guest_house (gîte@fr)
v_hostel (auberge de jeunesse@fr)
v_hotel (hôtel@fr)
v_information (office de tourisme@fr)
v_motel (motel@fr)
v_museum (musée@fr)
v_theme_park (parc d'attractions@fr)
v_zoo (parc zoologique@fr)

### k_highway (Routes)

v_bus_stop (arrêt de bus@fr)
v_cycleway (aménagement cyclable@fr)
v_motorway_link (échangeur autoroutier@fr)
v_motorway (autoroute@fr)
v_path (sentier@fr)
v_pedestrian (rue piétonne@fr)
v_primary (route nationale@fr)
v_raceway (circuit automobile@fr)
v_residential (rue résidentielle@fr)
v_secondary (route départementale@fr)
v_steps (escaliers@fr)
v_tertiary (route communal@fr)
v_trunk (voie rapide@fr)

### k_waterway (Cours d’eau)

v_canal (canal artificiel@fr)
v_dam ( barrage@fr)
v_ditch (fossé@fr)
v_drain (égout@fr)
v_riverbank (berge@fr)
v_river (fleuve@fr)
v_river (rivière@fr)
v_stream (ruisseau@fr)
v_waterfall (chute d'eau@fr)

### k_railway (Chemins de fer)

v_abandoned (ancienne voie de chemin de fer@fr)
v_crossing (passage piéton@fr)
v_disused (portion de voie ferrée désaffectée@fr)
v_halt (petite gare de train@fr)
v_level_crossing (passage à niveau@fr)
v_platform (quai@fr)
v_spur (voie ferrée de service@fr)
v_station (gare ferroviaire@fr)
v_tram_stop (arrêt de tramway@fr)

### k_man_made (Edifices)

```
v_pier (jetée@fr)
v_pipeline (aqueduc@fr)
v_pipeline (oléoduc@fr)
v_pipeline (gazoduc@fr)
v_tower (antenne@fr)
v_wastewater_plant (traitement de l'eau@fr)
v_water_tower (Château d'eau@fr)
v_works (Usine@fr)
```

### k_route (Itinéraires)

```
v_bus (ligne de bus@fr)
v_ferry (ligne de ferry@fr)
v_hiking (sentier de randonnée@fr)
v_mtb (itineraire de randonnée en vtt@fr)
v_railway (itineraire en reseau ferré@fr)
v_ski (piste de ski@fr)
v_train (ligne de train@fr)
v_tram (ligne de tramway@fr)
```

### k_historic (Patrimoine)

```
v_archaeological_site (site archéologique@fr)
v_boundary_stone (borne frontière@fr)
v_castle (château@fr)
v_memorial (commémoration@fr)
v_monument (monument@fr)
v_ruins (ruines@fr)
v_wayside_shrine (tombeau historique@fr)
```

### k_power (Energie)

```
v_generator (centrale électrique@fr)
v_line (câbles aériens à haute-tension@fr)
v_minor_line (câbles aériens à basse tension@fr)
v_pole (poteau de support de câbles à basse tension@fr)
v_pole (poteau de support de câbles à moyenne tension@fr)
v_sub_station (transformateur@fr)
v_transformer (transformateur électrique@fr
```

### k_military (Défense)

```
v_airfield (Aérodrome@fr)
v_barracks (Caserne@fr)
v_bunker (bunker@fr)
v_danger_area (zone de tir@fr)
v_naval_base (base navale@fr)
v_range (stand de tir@)
```

### k_aeroway (Aviation)

```
v_aerodrome (aéroport@fr)
v_gate (porte d'embarquement@fr)
v_helipad (héliport@fr)
v_runway (piste d'atterrissage@fr)
v_taxiway (voie de circulation@fr)
v_terminal (terminal aéroportuaire@fr)
```

### k_boundary (Frontières)

```
v_administrative (frontière@fr)
v_maritime (frontière maritime@fr)
v_national_park (parc national@fr)
v_protected_area (Aire protégée@fr)
```

### k_aerialway (Transports par câble)

v_cable_car (remontée mécanique@fr)
v_chair_lift (télésiège@fr)
v_drag_lift (téléski@fr)
v_gondola (télécabine@fr)

### k_power_source (Centrales d'électricité)

v_hydro (centrale hydroélectrique@fr)
v_nuclear (centrale nucléaire@fr)
v_photovoltaic (centrale solaire photovoltaïque@fr)
v_wind (centrale éolienne@fr)

### k_landuse (Utilisation du sol)

v_forest (forêt@fr)

### k_station (Stations de transport)

v_subway (station de métro@fr)

# Bibliographie


Abadie, N, 2012. *Intégration des bases de données à partir de la formalisation de leurs spécifications*. Université Paris-Est Marne-la-Vallée.

Abadie, N et Mustière, S, 2010. Constitution et exploitation d'une taxonomie géographique à partir des spécifications de bases de données. *Revue Internationale de Géomatique*.

Abbas, I., 1994. *Bases de données vectorielles et erreur cartographique. Problèmes posés par le contrôle ponctuel; une méthode alternative fondée sur la distance de Hausdorff*. Université Paris 7.

Adams B, Li L, Raubal M, Goodchild M F, 2010. A general framework for conflation. In R Purves & R. Weibel, eds. *in Proceedings of the 6th International Conference, GIScience 2010, Extended Abstracts Volume. R*. Zurich, Switzerland, pp. 1–5. Available at: http://www.giscience2010.org/pdfs/paper_211.pdf [Accessed November 23, 2012].

Antoniou, V., 2011. *User generated spatial content: an analysis of the phenomenon and its challenges for mapping agencies*. University College London. Available at: http://discovery.ucl.ac.uk/1318053/ [Accessed November 14, 2012].

Arkin, E, Chew, L, Huttenlocher, D, Kedem, K, Mitchell, J, 1991. An Efficiently Computable Metric for Comparing Polygonal Shapes. *in IEEE trans. On Pattern Recognition and Machine Intelligence*, 13(3), pp.209–216.

Atemezing, G. et Troncy, R., 2012. Vers une meilleure interopérabilité des données géographiques françaises sur le Web de données. In *IC 2012, 23èmes Journées Francophones d'Ingénierie des Connaissances*. Paris, France. Available at: http://www.eurecom.fr/publication/3717.

Badard, T., 2000. *Propagation des mises à jour dans les bases de données géographiques multi-représentations par analyse des changements géographiques*. Université Paris-Est Marne-la-Vallée.

Baglatzi, A., Kokla, M. et Kavouras, M., 2012. Semantifying OpenStreetMap. In D. Kolas et al., eds. *in Proceedings of the 5th International Terra Cognita Workshop 2012 In Conjunction with the 11th International Semantic Web Conference*. Boston, USA: CEUR. Available at: http://www.strabon.di.uoa.gr/terracognita/papers/proceedings.pdf.

Ballatore, A., Wilson, D.C. et Bertolotto, M., 2012. A Survey of Volunteered Open Ontologies in the Semantic Geospatial Web. In *Advanced Techniques in Web Intelligence - 3: Quality-based Information Retrieval*. Springer.

Bejaoui, L, Pinet, F, M, Schneider, Y, Bédard, 2010. OCL for formal modelling of topological constraints involving regions with broad boundaries. *GeoInformatica*, pp.353–378.

Bel Hadj Ali, A., 2001. *Qualité géométrique des entités géographiques surfaciques: application à l'appariement et définition d'une typologie des écarts géométriques*, la-Vall{é}e : 2001. Available at: http://books.google.fr/books?id=NT1KXwAACAAJ.

Bishr, M et Kuhn, W, 2007. Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In *Proceedings of AGILE*. pp. 365–387.

Bishr, M et J, Krzysztof, 2010. Can we trust information?-the case of volunteered geographic information. *Towards Digital Earth Search Discover …*. Available at: http://helios.geog.ucsb.edu/~jano/DE2010qp.pdf [Accessed November 14, 2012].

Bizer, C, Lehmann, J, Kobilarov, G, Auer, S, Becker, C, Cyganiak, R, Hellmann, S, 2009. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), pp.154–165. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1570826809000225.

Borges, K.A.V., Clodoveu, D.A. et Laender, A.H.F., 2002. Integrity Constraints in Spatial Databases. In J. H. Doorn & R. L. C, eds. *Database Integrity*. Hershey, PA, USA: IGI Publishing, pp. 144–171.

Bruns, H.T. et Egenhofer, Max J, 1996. Similarity of Spatial Scenes. In *7 th Symposium on Spatial Data Handling*. pp. 31–42.

Bucher, B, Falquet, G, Clementini, E, Sester, M, 2012. Towards a typology of spatial relations and properties for urban applications. In T. Leduc, G. Moreau, & R. Billen, eds. *Usage, Usability, and Utility of 3D City Models – European COST Action TU0801*. Nantes, France.

Bucher, B, 2009. *Vers la diffusion en ligne d'information géographique sur mesure*. Université Paris-Est, Marne-la-Vallée.

Bucher, B, 2011. L'information géographique sur internet: qualité des contenus, adaptation de la carte au contexte. In C. Boucher, ed. *actes de La Journée Scientifique du Bureau des longitudes, "La Nouvelle Géographie."*

Budhathoki, N., 2010. *Participants' motivations to contribute geographic information in an online community*. University of Illinois, IL, USA. Available at: https://www.ideals.illinois.edu/handle/2142/16956.

Bédard, Y, Larrivée, S, Proulx, MJ, Nadeau, M, 2004. Modeling Geospatial Data-bases with Plug-Ins for Visual Languages: A Pragmatic Approach and the Impacts of 16 Years of Research and Experimentation on Perceptory. In *in Proceedings of the Conceptual Modeling for Geographic Information Systems Workshop*. Springer LNCS 3289, pp. 17–30.

Carter, R. et Frith, C.D., 1998. *Mapping the Mind*, University of Calif. Press. Available at: http://books.google.fr/books?id=W6taDXzWMr0C.

Casado, M.L., 2006. Some Basic Mathematical Constraints for the Geometric Conflation Problem. In *in Proceedings of the 7th International Symposium on Spatial Accuracy*

*Assessment in Natural Resources and ASPRS 2010 Annual Conference San Diego, California*. San Diego, California, pp. 264–274.

Chen, C.-C. et Knoblock, C.A., 2008. Conflation of Geospatial Data. In S. Shekhar & H. Xiong, eds. *Encyclopedia of GIS*. Springer, pp. 133–140. Available at: http://dblp.uni-trier.de/db/reference/gis/gis2008.html#ChenK08.

Chipofya, M., Wang, J. et Schwering, A., 2011. Towards Cognitively Plausible Spatial Representations for Sketch Map Alignment. In Max Egenhofer et al., eds. *In Proceedings of the 10th international conference on Spatial information theory (COSIT'11)*. Berlin, Heidelberg: Springer-Verlag, pp. 20–39.

Claramunt, C., 2000. Extending Ladkin's algebra on non-convex intervals towards an algebra on union-of regions. In *Proceedings of the 8th ACM international symposium on Advances in geographic information systems*. New York, NY, USA: ACM, pp. 9–14. Available at: http://doi.acm.org/10.1145/355274.355276.

Cobb M A, Chung M J, Foley III H, Petry F E, Shaw K B, Miller, H V, 1998. A Rule-based Approach for the Conflation of Attributed Vector Data. *Geoinformatica*, 2(1), pp.7–35. Available at: http://dx.doi.org/10.1023/A:1009788905049.

Cockcroft, S., 1997. A Taxonomy of Spatial Data Integrity Constraints. , 343, pp.327–343.

Codescu, M, Horsinka, G, Kutz, O, Mossakowski, T, Rau, R, 2011. OSMonto - An Ontology of OpenStreetMap Tags. In *In State of the map Europe (SOTM-EU)*.

Coleman, D., Georgiadou, Y. et Labonte, J., 2009. Volunteered Geographic Information: the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, 4, pp.332–358. Available at: http://drupal.gsdi.org/gsdiconf/gsdi11/papers/pdf/279.pdf [Accessed November 21, 2012].

Deakin, A., 1996. Landmarks as Navigational Aids on Street Maps. *Cartography and Geographic Information Systems*, 23(1), pp.21–36.

Dempster, A.P., 1967. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2).

Deng, M., Chen, X.Y. et LI, Z., 2005. A Generalized Hausdorff Distance for Spatial Objects in GIS. In *International Archives of Photogrammetry and Remote Sensing*. Wales, UK, pp. 10–15.

Devillers, R., Bédard, Y. et Gervais, M., 2004. Indicateurs de qualité pour réduire les risques de mauvaise utilisation des données géospatiales. *Revue Internationale de Géomatique*, 14(1), pp.35–57.

Devillers, R. et Jeansoulin, R., 2006. *Fundamentals of Spatial Data Quality* ISTE., Newport Beach, CA.

Devogele, T, 2002. A new Merging process for data integration based on the discrete Fréchet distance. In D. Richardson & P. van Oosterom, eds. *in Proceedings of the 10th International Symposium on Spatial Data Handling (SDH)*. Ottawa, Canada, pp. 167–181.

Devogele, T, Parent, C. et Spaccapietra, S., 1998. On Spatial Database Integration. *International Journal of Geographical Information Science*, 12(4), pp.335–352. Available at: http://dblp.uni-trier.de/db/journals/gis/gis12.html#DevogelePS98.

Duchêne, C, 2004. *Généralisation par agents communicants : Le modèle CARTACOM. Application aux données topographiques en zone rurale*. Université Paris 6.

Duchêne, C, Bard, S, Barillot, X, Ruas, A, Trévisan, J, Holzapfel, F, 2003. Quantitative and qualitative description of building orientation. In *6th ICA Workshop on Generalisation and Multiple Representation, 28-30 April, Paris (France)*. Available at: http://bibliosr.ign.fr/Publications/2003/Duchene03d.

Duchêne, C, Ruas, A et Cambier, C., 2012. The CARTACOM model: transforming cartographic features into communicating agents for cartographic generalisation. *International Journal of Geographical Information Science*, 26(9), pp.1533–1562. Available at: http://bibliosr.ign.fr/Publications/2012/Duchene12.

Duckham, M, Lingham, J, Mason, K, Worboys, M, 2006. Qualitative reasoning about consistency in geographic information. *Information Sciences*, 176(6), pp.601–627. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0020025505002100 [Accessed September 28, 2012].

Edwards, W K, Mynatt, E D, Petersen, K, Spreitzer, M J, Terry, D B, Theimer, M M, 1997. Designing and implementing asynchronous collaborative applications with Bayou. In *Proceedings of the 10th annual ACM symposium on User interface software and technology*. New York, NY, USA: ACM, pp. 119–128. Available at: http://doi.acm.org/10.1145/263407.263530.

Egenhofer, M et Herring, J., 1991. *Categorizing Binary Topological Relationships Between Regions, Lines, and Points in Geographic Databases*, Orono, {ME}: Department of Surveying Engineering, University of Maine.

Egenhofer, M J et Franzosa, R., 1991. Point-set topological spatial relations. *International Journal of Geographic Information Systems*, 5(2), pp.161–174.

Egenhofer, M et Mark, D., 1995. Naive Geography. In A. U. Frank & W Kuhn, eds. *Spatial Information Theory: A Theoretical Basis for GIS, Lecture Notes in Computer Sciences, Springer-Verlag,*. Berlin, pp. 1–15.

Ellis, C.A. et Gibbs, S.J., 1989. Concurrency Control in Groupware Systems. In J. Clifford, B. G. Lindsay, & D. Maier, eds. *SIGMOD Conference*. ACM Press, pp. 399–407. Available at: http://dblp.uni-trier.de/db/conf/sigmod/sigmod89.html#EllisG89.

Elmasri, R. et Navathe, S., 1994. *Fundamentals of Database Systems*, Benjamin/Cummings.

Euzenat, J. et Shvaiko, P., 2007. *Ontology Matching*, Springer.

Exel, M.V., Dias, E. et Fruijtier, S., 2010. The impact of crowdsourcing on spatial data quality indicators. In *Proceedings of the 6th GIScience International Conference on Geographic Information Science*. Zurich, pp. 1–4.

Fielding, R.T., 2000. *REST: Architectural Styles and the Design of Network-based Software Architectures*. University of California, Irvine. Available at: http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

Fisher, P., Comber, A. et Wadsworth, R., 2009. What's in a Name? Semantics, Standards and Data Quality. In R. Devillers & H. Goodchild, eds. *Spatial Data Quality : From Process to Decisions*. p. 226.

Flanagin, A.J. et Metzger, M.J., 2008. The credibility of volunteered geographic information. *GeoJournal*, 72(3-4), pp.137–148. Available at: http://www.springerlink.com/index/10.1007/s10708-008-9188-y [Accessed July 22, 2012].

Frank, A, 2001. Tiers of Ontology and Consistency Constraints in Geographic Information Systems. *International Journal of Geographic Information Science*, 15(7), pp.667–678.

Gangemi, A, Guarino, N, Masolo, C, Oltramari, A, Schneider, L, 2002. Sweetening Ontologies with DOLCE. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*. London, UK, UK: Springer-Verlag, pp. 166–181. Available at: http://dl.acm.org/citation.cfm?id=645362.650863.

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

Gillman, D.W., 1985. Triangulations for Rubber-Sheeting. In *in Proceedings of AUTOCARTO 7*. Washington D.C., USA, pp. 191–199.

Girres, J.-F. et Touya, G, 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4), pp.435–459. Available at: http://doi.wiley.com/10.1111/j.1467-9671.2010.01203.x [Accessed September 24, 2012].

Goodchild, M, 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp.211–221. Available at: http://dx.doi.org/10.1007/s10708-007-9111-y.

Goodchild, M, 2009. NeoGeography and the nature of geographic expertise. *Journal of Location Based Services*, 3(2), pp.82–96. Available at: http://www.tandfonline.com/doi/abs/10.1080/17489720902950374 [Accessed July 15, 2012].

Goodchild, M F et Hunter, G.J., 1997. A Simple Positional Accuracy Measure for Linear Features. *International Journal of Geographical Information Science*, 11(3), pp.299–306. Available at: http://dblp.uni-trier.de/db/journals/gis/gis11.html#GoodchildH97.

Grosso, E., Perret, J. et Brasebin, M., 2012. GEOXYGENE: an Interoperable Platform for Geographical Application Development. In *Innovative Software Development in Gis*. John Wiley & Sons, pp. 67–90. Available at: http://bibliosr.ign.fr/Publications/2012/Grosso12.

Hadzilacos, T. et Tryfona, N., 1992. A Model for Expressing Topological Integrity Constraints in Geographic Databases. In A U Frank, I. Campari, & U. Formentini, eds. *Proceedings of the International Conference GIS - From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning on Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*. London, UK: Springer-Verlag, pp. 252–268.

Haklay, M. et Basiouka, S., 2010. How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. *Cartographic Journal, …*, pp.1–13. Available at: http://www.ingentaconnect.com/content/maney/caj/2010/00000047/00000004/art00005 [Accessed November 14, 2012].

Haklay, M.M., 2010. How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England. *Environment and Planning B*, 37(4), pp.682 – 703.

Haunert, J.-H., 2005. Link based conflation of geographic datasets. In *Proceedings of the 8th ICA Workshop on Generalisation and Multiple Representation, July 7--8, 2005, A Coruna, Spain*.

Hecht, B. et Raubal, M., 2008. GeoSR: Geographically Explore Semantic Relations in World Knowledge. In Lars Bernard, A. Friis-Christensen, & H. Pundt, eds. *AGILE Conf.* Springer, pp. 95–113. Available at: http://dblp.uni-trier.de/db/conf/agile/agile2008.html#HechtR08.

Horrocks, I, Patel-Schneider, P F, Boley, H, Tabet, S, Grosof, B, Dean, M, 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C Member Submission 21 May 2004. Available at: http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/.

Huberman, A M, Miles, M B, Rispal, M H, Bonniol, J J, 2003. *Analyse des données qualitatives*, De Boeck Supérieur. Available at: http://books.google.fr/books?id=AQHRyJ1AiPEC.

IGN, 2011a. API-REST pour RIPart V 0.1 - Manuel utilisateur, Projet Echanges.

IGN, 2011b. BDUni V1.1 Grande Échelle Spécifications de qualité.

IGN, 2009a. Contrôle Qualité BD Uni Départemental et Unification du Bâti sur le département 14 (Calvados).

IGN, 2009b. Contrôle Qualité de la BDUni Manuel Opérateur.

IGN, 2006. Manuel opérateur GCVS pour le collecteur.

IGN, 2011c. Spécifications de la BDTopo® de l'IGN. Available at: http://professionnels.ign.fr/sites/default/files/DC_BDTOPO_2_1.pdf.

ISO, 2005. ISO 19109: Geographic information - Rules for application schema. , (19109).

ISO, 2002. ISO 19113: Geographic information - Quality principles.

ISO, 2003. ISO 19115: Geographic information - Metadata. , (19115).

Jaara, K., Duchêne, C et Ruas, A, 2012. A model for preserving the consistency between topographic and thematic layers throughout data migration. In *in Proceedings of the 5th International Symposium on Spatial Data Handling (SDH'12)*. Bonn, Germany.

Kergosien, E, Kamel, M, Sallaberry, C, Bessagnet, M-N, Aussenac-Gilles, N, Gaio, M, 2010. Construction et enrichissement automatique d'ontologie à partir de ressources externes. *CoRR*, abs/1002.0. Available at: http://dblp.uni-trier.de/db/journals/corr/corr1002.html#abs-1002-0239.

Keßler, C, Johannes, T. et Janowicz, K, 2011. Tracking Editing Processes in Volunteered Geographic Information: The Case of OpenStreetMap. In *Proceedings of COSIT'11 Workshop IOPE*. ACM. Available at: http://dx.doi.org/10.1145/1653771.1653787.

Kittur, A, Suh, B, Pendleton, B A, Chi, E H, 2007. He Says, She Says: Conflict and Coordination in Wikipedia +. , pp.453–462.

Krötzsch, M, Vrandečić, D, Völkel, M, Haller, H, Studer, R, 2007. Semantic Wikipedia. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4), pp.251–261. Available at: http://dx.doi.org/10.1016/j.websem.2007.09.001.

Kuhn, W, 2007. Volunteered geographic information and GIScience. *NCGIA, UC Santa Barbara*. Available at: http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Kuhn_paper.pdf [Accessed November 14, 2012].

Lamport, L., 1978. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM*, 21(7), pp.558–565. Available at: http://dx.doi.org/10.1145/359545.359563.

Laurini, R et Thompson, D., 1992. *Fundamentals of Spatial Information Systems*, Academic Press, Inc.

Li, L. et Goodchild, M F, 2011. An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion*, 2(4), pp.309–328. Available at: http://www.tandfonline.com/doi/abs/10.1080/19479832.2011.577458.

Limpens, F., Gandon, F. et Buffa, M., 2009. Sémantique des folksonomies: structuration collaborative et assistée. In F. L. Gandon, ed. *Actes d'IC*. PUG, pp. 37–48. Available at: http://dblp.uni-trier.de/db/conf/f-ic/ic2009.html#LimpensGB09.

Longley, P A, Goodchild, M F, Maguire, D J, Rhind, D W, 2005. *Geographic Information Systems and Science*, John Wiley & Sons. Available at: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0470870001.

Mackaness, W., Beard, K. et Buttenfield, B., 1994. *Selected Annotated Bibliography on Visualization of the Quality of Spatial Information*,

Mark, D.M. et Frank, A U, 1989. Concepts of Space and Spatial Language. In *in Proceedings of the 9th International Symposium on Computer-Assisted Cartography (AUTOCARTO)*. Baltimore, Maryland, USA, pp. 538–556.

Martin, S., 2011. *Edition collaborative des documents semi-structurés*. Université de Provence - Aix-Marseille 1.

Massiot, L., Abadie, N et Bucher, B, 2011. Mining user generated Web content to characterise geographic features beyond topographical aspects. In *in Proceedings of the 25th International Cartographic Conference (ICC'11)*. Paris, France.

Meier, W., 2003. eXist: An open source native XML database. In *Revised Papers from the NODe 2002 Web and Database-Related Workshops on Web, Web-Services, and Database Systems*. London, UK: Springer-Verlag, pp. 169–183. Available at: http://www.springerlink.com/index/DV46DPQMCND4E1F3.pdf [Accessed October 24, 2012].

Mennis, J.L., Peuquet, D.J. et Qian, L., 2000. A conceptual framework for incorporating cognitive principles into geographical database representation. *International Journal of Geographical Information Science*, 14(6), pp.501–520. Available at: http://dx.doi.org/10.1080/136588100415710.

Michaux, J, Blanc, X, Shapiro, M, Sutra, P, 2011. A semantically rich approach for collaborative model edition. In *SAC*. ACM, pp. 1470–1475.

Mihalcea, R., 2007. Using Wikipedia for Automatic Word Sense Disambiguation. In *North American Chapter of the Association for Computational Linguistics (NAACL 2007)*.

Miller, G., 1995. WordNet: a Lexical Database for English. *Communications of the ACM*, Volume 38(Number 11).

Miron, A D, Gensel, J, Villanova-Olivier, M, Martin, H, 2007. Relations spatiales qualitatives dans les ontologies géographiques avec ONTOAST. In *Colloque Int. de Géomatique et d'Analyse Spatiale (SAGEO 2007)*. Clermont-Ferrand, France.

Molli, P, Oster, G, Skaf-Molli, H, Imine, A, 2003. Using the transformational approach to build a safe and generic data synchronizer. *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work - GROUP '03*, p.212. Available at: http://portal.acm.org/citation.cfm?doid=958160.958194.

Mooney, P. et Corcoran, P., 2011. Accessing the history of objects in OpenStreetMap. In *Proceedings of the 14h AGILE International …*. Utrecht, The Netherlands, pp. 2–4. Available at: http://plone.itc.nl/agile_old/Conference/2011-utrecht/contents/pdf/posters/p_115.pdf [Accessed November 21, 2012].

Mustière, S et Devogele, T, 2008. Matching Networks with Different Levels of Detail. *Geoinformatica*, 12(4), pp.435–453. Available at: http://dx.doi.org/10.1007/s10707-007-0040-1.

Mäs, S, Wang, F. et Reinhardt, W, 2005. Using Ontologies for Integrity Constraint Definition. In *In Proceedings of the 4th International Symposium On Spatial Data Quality (ISSDQ'05)*. ACM, pp. 304–313.

Mäs, S et Reinhardt, W, 2009. Categories of Geospatial and Temporal Integrity Constraints. *Advanced Geographic Information Systems & Web Services, International Conference on*, 0, pp.146–151. Available at: http://doi.ieeecomputersociety.org/10.1109/GEOWS.2009.12.

Mülligann, C, Janowicz, K, Ye, M, Lee, W-C, 2011. Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information. In Max J Egenhofer et al., eds. *Spatial Information Theory - COSIT*. Springer, pp. 350–370. Available at: http://dblp.uni-trier.de/db/conf/cosit/cosit2011.html#MulligannJYL11.

Nastase, V, Strube, M, Boerschinger, B, Zirn, C, Elghafari, A, 2010. WikiNet: A Very Large Scale Multi-Lingual Concept Network. In N. C. (Conference Chair) et al., eds. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

OGC, 2008. *OWS-5 Conflation Engineering Report*,

OSM, 2012a. API v0.6 : Changesets. *API v0.6 : Changesets*. Available at: http://wiki.openstreetmap.org/wiki/API_v0.6#Changesets [Accessed August 31, 2012].

OSM, 2012b. Beginners Guide 1.1 - General Tips. *Beginners Guide 1.1 - General Tips*. Available at: http://wiki.openstreetmap.org/wiki/Beginners_Guide_1.1 [Accessed August 28, 2012].

OSM, 2012c. FR:PlaceMaker. Available at: http://wiki.openstreetmap.org/wiki/FR:PlaceMaker [Accessed August 27, 2012].

OSM, 2012d. Osmosis. *Osmosis*. Available at: http://wiki.openstreetmap.org/wiki/Osmosis [Accessed August 31, 2012].

OSM, 2012e. Quality assurance. Available at: http://wiki.openstreetmap.org/wiki/Quality_assurance [Accessed August 29, 2012].

OSM, 2012f. Relations are not categories. *Relations are not categories*. Available at: http://wiki.openstreetmap.org/wiki/FR:Relations/Relations_are_not_Categories [Accessed August 28, 2012].

OSM, 2012g. WikiProject France/Cadastre/Import automatique des bâtiments. *WikiProject France/Cadastre/Import automatique des bâtiments*. Available at: http://wiki.openstreetmap.org/wiki/WikiProject_France/Cadastre/Import_automatique_des_bâtiments [Accessed August 28, 2012].

OSM, 2012h. WikiProject France/Corine Land Cover. *WikiProject France/Corine Land Cover*. Available at: http://wiki.openstreetmap.org/wiki/WikiProject_France/Corine_Land_Cover [Accessed August 27, 2012].

Olteanu, A.-M., 2008. *Appariement de données spatiales par prise en compte de connaissances imprécises*. Université de Marne-La-Vallée.

Oster, G, Urso, P, Molli, P, Imine, A, 2006. Data Consistency for P2P Collaborative Editing. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. New York, NY, USA: ACM Press, pp. 259–268. Available at: http://dx.doi.org/10.1145/1180875.1180916.

Oster, G, Skaf-Molli, H, Molli, P, Naja-Jazzar, H, 2007. Supporting Collaborative Writing of XML Documents. In *9th International Conference on Enterprise Information Systems - ICEIS 2007*. Funchal, Madeira, Portugal, pp. 335–341. Available at: http://hal.inria.fr/inria-00139704.

Overell, S., Ruger, S. et Magalhaes, J., 2006. Place disambiguation with co-occurrence models. In A. Nardi, C. Peters, & J. L. Vicedo, eds. *CLEF 2006 Workshop, Working notes*. Available at: http://mmir.doc.ic.ac.uk/www-pub/geoclef06.pdf.

Parent, C, Spaccapietra, S, Zimanyi, E, Donini, P, Plazanet, C, Vangenot, C, 1998. Modeling spatial data in the MADS conceptual model. In *in Proceedings of the 8th International Symposium on Spatial Data Handling, SDH'98*. pp. 138–150.

Pinet, F, Duboisset, M. et Schneider, M., 2009. Modélisation de contraintes d'intégrité spatiales avec OCL. *Revue Internationale de Géomatique*, 19(1), pp.93–122. Available at: http://dblp.uni-trier.de/db/journals/rig/rig19.html#PinetDS09.

Preguiça, N.M., Shapiro, M et Matheson, C., 2003. Semantics-Based Reconciliation for Collaborative and Mobile Environments. In R. Meersman, Z. Tari, & D. C. Schmidt, eds. *CoopIS/DOA/ODBASE*. Springer, pp. 38–55. Available at: http://dblp.uni-trier.de/db/conf/coopis/coopis2003.html#PreguicaSM03.

Purves, R, Edwardes, A.J. et Wood, J., 2011. Describing place through user generated content. *First Monday*, 16(9). Available at: http://dblp.uni-trier.de/db/journals/firstmonday/firstmonday16.html#PurvesEW11.

Ramm, F., Topf, J. etChilton, S., 2010. *OpenStreetMap: Using and Enhancing the Free Map of the World*, Uit Cambridge Limited. Available at: http://books.google.fr/books?id=AnCNQQAACAAJ.

Raper, J., 1996. Unsolved problems of spatial representation. In *in Proceedings of SDH'96*. pp. 14.1–14.11.

Raymond, E.S., 1999. *The Cathedral and the Bazaar* 1st ed. T. O'Reilly, ed., Sebastopol, CA, USA: O'Reilly & Associates, Inc.

Renard, J, Gaffuri, J, Duchêne, C, Touya, G, 2011. Automated generalisation results using the agent-based platform CartAGen. In *25th International Cartographic Conference (ICC'11)*. Available at: http://bibliosr.ign.fr/Publications/2011/Renard11.

Renz, J. et Nebel, B., 1998. Spatial reasoning with topological information. *Spatial Cognition*. Available at: http://www.springerlink.com/index/7G7JE0BDD4PTG2FM.pdf [Accessed December 21, 2012].

Ressler, J., Freese, E. et Boaten, V., 2009. Semantic Method of Conflation Matching problem. In *International Semantic Web Conference - Terra Cognita Workshop.* pp. 1–15.

Roche, S. et Feick, R., 2012. Wiki-place: Building place-based GIS from VGI. In *position paper from VGI Worshop Role of Volunteered Geographic Information in Advancing Science: Quality and Credibility, in conjunction with GIScience 2012*. Columbus, Ohio, USA.

Roick, O., Hagenauer, J. et Zipf, A., 2011. OSMatrix–grid-based analysis and visualization of OpenStreetMap. In *Proceedings of the 1st State of the Map EU*. Vienne, Autriche. Available at: http://koenigstuhl.geog.uni-heidelberg.de/publications/2011/Roick/Roick_2011_SotM.pdf [Accessed November 16, 2012].

Roussey, C. et Pinet, F, 2010. DL based automated consistency checking of spatial relationships. In *Conférence internationale de Géomatique et Analyse Spatiale SAGEO'10*. pp. 306–320. Available at: http://liris.cnrs.fr/publis/?id=4908.

Ruas, A et Mackaness, W., 1997. Strategies for urban map generalisation. In *in Proceedings of the 18th International Cartographic Conference (ICC'97)*. Stockholm, Sweden.

Ruas, A et Plazanet, C, 1996. Strategies for automated map generalisation. In *7th International Symposium on Spatial Data Handling (SDH'96)*. Delft, Netherlands, pp. 319–336.

Ruas, A, 1999. *Modèle de généralisation de données géographiques à base de contraintes et d'autonomie*. Université de Marne-la-Vallée, Laboratoire COGIT, Institut Géographique National.

Ruiz, J J, Ariza, F J, Urena, M A, Blazquez, E B, 2011. Digital map conflation: a review of the process and a proposal for classification. *IJGIS*, 25(9), pp.1439–1466.

Saalfeld, A.J., 1993. *Conflation: automated map compilation*. College Park, MD, USA: University of Maryland at College Park.

Scheider, S, Keßler, C, Ortmann, J, Devaraju, A, Trame, J, Kauppinen, T, Kuhn, W, 2011. Semantic referencing of geosensor data and volunteered geographic information. In N. Ashish & A. P. Sheth, eds. *Geospatial Semantics and the Semantic Web*. Springer, pp. 27–59. Available at: http://dblp.uni-trier.de/db/conf/swb/geo2011.html#ScheiderKODTKK11.

Servigne, S, Ubeda, T, Puricelli, A, Laurini, R, 2000. A Methodology for Spatial Consistency Improvement of Geographic Databases. *Geoinformatica*, 4(1), pp.7–34. Available at: http://dx.doi.org/10.1023/A:1009824308542.

Shafer, G., 1976. *A Mathematical Theory of Evidence*, Princeton: Princeton University Press.

Sheeren, D., Mustière, S et Zucker, J.-D., 2004. How to Integrate Heterogeneous Spatial Databases in a Consistent Way? In *Advances in Databases and Information Systems*. pp. 364–378. Available at: http://www.springerlink.com/content/xl1k535x5v06p101.

Stankutė, S. et Asche, H., 2011. Improvement of spatial data quality using the data conflation. In *Proceedings of the 2011 international conference on Computational science and its applications - Volume Part I*. Berlin, Heidelberg: Springer-Verlag, pp. 492–500. Available at: http://dl.acm.org/citation.cfm?id=2029487.2029523.

Sterling, L. etShapiro, E., 1994. *The Art of Prolog*, MIT Press.

Taylor-Powell, E. et Renner, M., 2003. *Analyzing Qualitative Data*, University of Wisconsin--Extension, Cooperative Extension. Available at: http://books.google.fr/books?id=zuM7HAAACAAJ.

Touya, G, Coupé, A, Le Jollec, J, Dorie, O, Fuchs, F, 2012. Conflation Optimised by Least Squares to Maintain Geographic Shapes. *JOSIS*.

Touya, G, Duchêne, C et Ruas, A, 2010. Collaborative generalisation: formalisation of generalisation knowledge to orchestrate different cartographic generalisation processes. In *Proceedings of GIScience'10*. Berlin, Heidelberg: Springer-Verlag, pp. 264–278. Available at: http://dl.acm.org/citation.cfm?id=1887961.1887980.

Turner, B.A.J., 2006. *Introduction to Neogeography Introduction to Neogeography*,

Vasseur, B, Jeansoulin, R, Devillers, R, Frank, A, 2005. Evaluation de la qualité externe de l'information géographique: une approche ontologique. In R. Devillers & R. Jeansoulin, eds. *Qualité de l'information géographique: Traité Igat*. pp. 285–301.

Vauglin, F., 1997. *Modèles statistiques des imprécisions géométriques des objets géographiques linéaires*. Université de Marne-la-Vallée.

Viglino, J.-M., 2010. Intégration de mises à jour au travers de lots différentiels. In *Rencontres SIG la Lettre, ENSG*. Marne-la-Vallée, France.

Viglino, J.-M., 2011. Vers un système collaboratif pour la mise à jours de référentiels géographique. In *in Proceedings of the 25th International Cartographic Conference (ICC 2011)*. Paris, France.

Walter, V. et Fritsch, D., 1999. Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science*, 13(5), pp.445–473.

Wang, J., Mülligann, C. et Schwering, A., 2011. An Empirical Study on Relevant Aspects for Sketch Map Alignment. In S. C. M. Geertman, Wolfgang Reinhardt, & F. Toppen, eds. *The 14th AGILE International Conference on Geographic Information Science (AGILE 2011)*. Utrecht, The Netherlands: Springer, pp. 497–518. Available at: http://dblp.uni-trier.de/db/conf/agile/agile2011.html#WangMS11.

Ware, J.M. et Jones, C.B., 1998. Matching and Aligning Features in Overlayed Coverages. In Robert Laurini, K. Makki, & N. Pissinou, eds. *ACM-GIS*. ACM, pp. 28–33. Available at: http://dblp.uni-trier.de/db/conf/gis/gis98.html#WareJ98.

Watson, G.A., 2006. Computing Helmert transformations. *J. Comput. Appl. Math.*, 197(2), pp.387–394. Available at: http://dx.doi.org/10.1016/j.cam.2005.06.047.

Weiss, S., Urso, P et Molli, P, 2007. Wooki: a P2P Wiki-based Collaborative Writing Tool. In *Proceedings of WISE'07*. Springer, pp. 503–512.

Werder, S., 2009. Formalization of Spatial Constraints. In *in Proceedings of the 12th AGILE International Conference on Geographic Information Science (AGILE 2009)*. Hannover, Germany.

Wiemann, S. et Bernard, L, 2010. Conflation Services within Spatial Data Infrastructures. In *in proceedings of the 13th International Conference on Geographic Information Science (AGILE),*. Guimaraes, Portugal.

Wu, F. et Weld, D.S., 2008. Automatically refining the wikipedia infobox ontology. In *WWW*. pp. 635–644.

Yasseri, T, Sumi, R, Rung, A, Kornai, A, Kertész, J, 2012. Dynamics of conflicts in Wikipedia. *PloS one*, pp.1–36. Available at: http://dx.plos.org/10.1371/journal.pone.0038869 [Accessed December 21, 2012].

Yuan, S. et Tao, C., 1999. Development of Conflation Components. In *Proceedings of Geoinformatics'99*. pp. 1–13.

Zesch, T. et Gurevych, I., 2007. Analysis of the Wikipedia Category Graph for NLP Applications. , (April), pp.1–8.

Zesch, T. et Gurevych, I., 2010. Wisdom of crowds versus wisdom of linguists ? measuring the semantic relatedness of words. *Natural Language Engineering*, 16(01), pp.25–59. Available at: http://dx.doi.org/10.1017/S1351324909990167.

Zielstra, D. et Zipf, A., 2010. A comparative study of proprietary geodata and volunteered geographic information for germany. In *In 13th AGILE International Conference on Geographic Information Science*. pp. 1–15. Available at: http://agile2010.dsi.uminho.pt/pen/shortpapers_pdf/142_doc.pdf [Accessed November 14, 2012].

# Table des figures

# Table des tableaux

# Table des équations