Un nouveau méta-modèle pour rapprocher la folksonomie et l'ontologie d'OSM

Anthony Hombiat, Marlène Villanova-Oliver, Jérôme Gensel

Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
prenom.nom@imag.fr

RÉSUMÉ. *Depuis les années 2000, les technologies du Web permettent aux utilisateurs de prendre part à la production de données : les internautes du Web 2.0 sont les nouveaux capteurs de l'information. Du côté de l'Information Géographique affluent de nombreux jeux de données en provenance de plates-formes de cartographie participative telles qu'OpenStreetMap (OSM) qui a largement impulsé le phénomène de la Géographique Participative (VGI). La communauté OSM représente aujourd'hui plus de deux millions de contributeurs qui alimentent une base de données géospatiales ouverte dont l'objet est de capturer une représentation du territoire mondial. Les éléments cartographiques qui découlent de ce déluge de VGI sont caractérisés par des tags. Les tags permettent une catégorisation simple et rapide du contenu des plates-formes de crowdsourcing qui inondent la toile. Cette approche est cependant un obstacle majeur pour le partage et la réutilisation de ces grands volumes d'information. En effet, ces ensembles de tags, ou folksonomies, sont des modèles de données beaucoup moins expressifs que les ontologies. Dans cet article, nous proposons un méta-modèle pour rapprocher la folksonomie et l'ontologie OSM afin de mieux exploiter la sémantique des données qui en sont issues, tout en préservant la flexibilité intrinsèque à l'utilisation de tags.*

ABSTRACT. *Post-2000s web technologies have enabled users to engage in the information production process: Web 2.0 surfers are the new data sensors. Regarding Geographic Information (GI), large crowdsourced datasets emerge from the Volunteered Geographic Information (VGI) phenomenon through platforms such as OpenStreetMap (OSM). The latter involves more than two millions contributors who aim at mapping the world into an open geospatial database. This deluge of VGI consists of spatial features associated with tags describing their attributes which is typical of crowdsourced content categorization. However, this approach is also a major impediment to interoperability with other systems that could benefit from this huge amount of bottom-up data. Indeed, folksonomies are much less expressive data models than ontologies. In this paper, we address the issue of loose OSM metadata by proposing a model for collaborative ontology engineering in order to semantically lift the data while preserving the flexible nature of the activity of tagging.*

MOTS-CLÉS : *Information Géographique Volontaire (IGV), Données ouvertes, Web Sémantique, Ontologie*

KEYWORDS: *Volunteered Geographic Information (VGI), OpenData, Semantic Web, Ontology*

1. Introduction

Avec l'avènement des technologies du Web 2.0, les plates-formes comme wikipedia.org [1] et son analogue géospatial OpenStreetMap (OSM) [2] encouragent la production d'énormes quantités de contenus produits par les utilisateurs (*User Generated Content*, UGC). Alors que la barrière entre la production et la consommation de l'information tombe, A. Bruns invente le terme "*produsage*" pour parler du processus collaboratif et en continu de construction et d'extension de contenus existants dans le but d'une amélioration future (Bruns, 2006). Les *produsers* sont animés à la fois par des motivations constructives (altruisme, stimulation intellectuelle, expression personnelle, reconnaissance sociale, *etc*.) et nuisible (malice, calcul politique, intention criminelle, *etc*.) (Coleman *et al*., 2009). Au-delà de ces mobiles, la création d'UGC est stimulée par l'émergence de licences ouvertes. Parmi elles, les licences Creative Commons (CC) [3] essaient de trouver un équilibre entre le concept de "tous droits réservés" qui régit la propriété intellectuelle, et celui de "aucun droits réservés" que l'on retrouve dans le domaine public (Loenen, 2012). Cette nouvelle donne qui bouleverse le processus de production de l'information amène à sa suite nombre d'avantages mais soulève également de multiples problèmes.

Tout d'abord, du côté des avantages, la donnée citoyenne est le plus souvent volontaire, c'est-à-dire que toute personne équipée d'un accès à Internet peut potentiellement ajouter, modifier ou supprimer de l'information, ce qui a pour effet de générer des flux massifs de données ouvertes. De plus, cette information dite *crowdsourcée* est typiquement encadrée par les licences CC et donc libre de droits pour leur exploitation et réutilisation gratuite. D'autre part, le processus de collecte des données s'effectue en continu : les produsers peuvent alimenter le flux d'information à tout moment, garantissant par là même son actualisation. En ce qui concerne l'Information Géographique (IG), la donnée peut être collectée à un niveau de granularité très fin puisque les contributeurs ont tendance à décrire les zones dont ils sont familers (leur quartier, leur lieu de travail) à l'échelle de leurs déplacements quotidiens. Du même coup, la représentation qu'ils donnent de l'environnement avec lequel ils sont familier exprime une expérience terrain que les moyens limités d'une expertise (temps, ressources humaines) peinent à saisir. Enfin, les métadonnées sont, pour l'essentiel, associées aux données citoyennes par le biais de *tags* : ce sont des mots-clés en texte libre choisis par les utilisateurs dans le but de caractériser les resources issues de l'UGC. A titre d'exemple, pour décrire une école dans la base de données OSM, un contributeur doit d'abord dessiner ses limites géographiques sur la carte (c'est le niveau des instances, ou ABox en logique de description), puis il peut lui associer l'étiquette (*tag*) `ame-`

1. https://www.wikipedia.org/
2. http://www.openstreetmap.org/
3. http://creativecommons.org/licenses/

`nity=school` (c'est le niveau conceptuel, ou TBox en logique de description). Ces tags sous la forme `clé=valeur` sont décrits en texte libre, ce qui leur confère un caractère évolutif (Ho, Rajabifard, 2010).

Néanmoins, la donnée citoyenne ne va pas sans poser de problèmes de qualité parmi lesquels l'incomplétude (Neis *et al.*, 2011), le manque d'expertise (McCall, Minang, 2005) et le manque d'expressivité de son méta-modèle (Haklay, 2010). Dans cet article, nous nous concentrons sur ce dernier aspect. Les métadonnées d'OSM constituent une *folksonomie*, c'est-à-dire "le résultat de l'activité d'association libre d'étiquettes (*tags*) à des données ou objets (tout ce qui est identifié par une URL) dans le but personnel de les retrouver ultérieurement" (Wal, 2005). Cette méthode de catégorisation par étiquettage (*tagging*) est très flexible puisque l'utilisateur ne rencontre aucune contrainte lorsqu'il décrit du contenu avec des tags en texte libre. Pourtant, l'utilisation de foksonomies comporte de sérieux inconvénients tels que les coquilles, la redondance conceptuelle et le manque d'expressivité par rapport à d'autres types de classification tels que les thesauri (synonymie), les taxonomies (subsomption), ou encore les ontologies (logique du premier ordre) (Weller, 2007). Par ailleurs, l'usage des tags peut être ambigu : S. Golder identifie sept fonctions différentes des tags, chacune liée à une intention différente de l'utilisateur (Golder, 2006) : l'extraction d'information devient alors d'autant plus difficile. Dans cet article, nous proposons un méta-modèle pour la construction collaborative d'une ontologie des tags OSM dans le but d'améliorer la sémantique des données, tout en préservant la plasticité de l'activité de tagging. La suite de cet article s'articule de la manière suivante : la section 2 souligne les points forts et les faiblesses des efforts qui ont été faits pour structurer le méta-modèle d'OSM ; la section 3 présente notre méta-modèle et la section 4 résume nos contributions et donne des pistes vers d'autres travaux à réaliser dans la même direction.

2. État de l'art

Plusieurs travaux ont abordé les problèmes liés à la folksonomie d'OpenStreetMap. Dans la section suivante, nous dressons un aperçu de ces différentes propositions en mettant l'accent sur les points positifs et négatifs de chacune.

2.1. OSM wiki

Afin de décrire aussi précisément que possible les caractéristiques physiques du terrain, les contributeurs d'OSM doivent partager un vocabulaire commun. Dans l'optique de s'accorder sur le sens des tags, le projet OSM s'appuie sur un guide de bonnes pratiques sous la forme d'un wiki [4] qui regroupe des tags qui font consensus (2047 tags en 2013 d'après (Ballatore *et al.*, 2013)), ainsi que la description de l'utilisation qui doit en être faite dans le but de les désambiguïser. Mais surtout, comme n'importe

4. http://wiki.openstreetmap.org/wiki/Map_Features

quel wiki, les directives OSM sont établies par les contributeurs OSM et chaque tag choisi pour y figurer peut être soumis au cycle Contribution/Révocation/Discussion (*BOLD/Revert/Discuss, BRD*) [5] qui laisse place au débat : l'ajout d'un tag litigieux peut potentiellement être annulé via le système de gestion des versions propre au wiki. Le cycle BRD permet aux utilisateurs impliqués dans un contentieux de créer une page de discussion associée au tag concerné afin d'argumenter leurs choix et d'arriver, *in fine*, à un accord. Cette méthode répond parfaitement aux exigences de l'approche ascendante que défend le projet OSM. Cependant, même si les tags sélectionnés sont les plus fréquents dans la base de données OSM, ils ne représentent qu'une infime partie des 77 millions de tags [6] existants. Pire, aucune relation ne les lie, ce qui empêche tout raisonnement automatique sur le modèle dans son ensemble.

2.2. LinkedGeoData (LGD)

Les travaux de S. Auer comptent parmi les premières tentatives de construction d'une ontologie, LinkedGeoData (LGD) [7], des données (ABox) et métadonnées (TBox) d'OSM (Auer *et al*., 2009). Dans ce modèle, les tags sont classés en trois catégories, chacune correspondant à un modèle de conversion vers le langage de représentation des connaissances OWL [8] :

Attributs de classification Les attributs de classification sont des tags qui donnent des informations sur la nature des éléments spatiaux auxquels ils sont associés (ex. : `amenity=school`). Dans l'ontologie, à la fois la clé et la valeur sont représentées par des classes (`owl:Class`), la clé étant super-classe de la valeur (`rdfs:subClassOf`) ;

Attributs de description Les attributs de description sont des tags ayant pour valeur un ensemble de valeurs prédéfinies (ex. : `internet_access=wired/-wlan/terminal`). Ils sont convertis en propriétés objet (`owl:ObjectProperty`) dans l'ontologie ;

Attributs de données Les attributs de données sont des tags dont la valeur est textuelle ou de type primitif (ex. : `opening_hours=9am-5pm`). Ils sont convertis en propriétés primitives (`owl:DataProperty`) dans l'ontologie.

Néanmoins, cette caractérisation haut-niveau des tags est le reflet d'une expertise qui n'a pas été soumise à un processus démocratique qui est pourtant au cœur de la philosophie du projet OSM.

En fin de compte, l'ontologie LGD contient 500 classes, 50 propriétés objet et 15 000 propriétés primitives. De plus, les instances (c'est-à-dire les objets spatialisés

5. https://en.wikipedia.org/wiki/Wikipedia:BOLD,_revert,_discuss_cycle
6. http://wiki.openstreetmap.org/wiki/Taginfo/FAQ
7. http://wiki.openstreetmap.org/wiki/LinkedGeoData
8. *Web Ontology Language*, http://www.w3.org/2001/sw/wiki/OWL

dans la base de données OSM) sont interconnectées avec les entités DBpedia [9] pour lesquelles des valeurs de longitude et de latitude sont définies. La correspondance entre les deux bases de connaissances se fait par l'intermédiaire de la comparaison du nom, de la localisation et du type de ces entités. Toutefois, les classes dans l'ontologie (c'est-à-dire les clés et valeurs des tags) ne sont directement connectées à aucune autre source, ce qui freine considérablement l'interopérabilité des informations sur les tags.

2.3. OSMonto

L'ontologie OSMonto [10] (Codescu *et al.*, 2011) a été créée pour appuyer un outil de navigation orienté activités pour OSM [11]. La méthodologie qui sous-tend sa création est semblable à celle de l'ontologie LGD : les clés des tags OSM sont considérées comme les super-classes des valeurs auxquelles elles sont associées. Afin de prévenir l'ambiguïté dans le cas où elles sont les mêmes, les clés et valeurs sont respectivement préfixées par `k_` (ex. : `station=subway` donne `k_station`) et `v_` (ex. : `railway=station` donne `v_station`). Les dépendances entre tags sont également prises en considération : certains tags (ex. : `cuisine=seafood`) ne sont compatibles qu'avec d'autres tags bien spécifiques (ex. : `amenity=restaurant`). Dans ces cas, la clé du tag dépendant est convertie en une propriété objet (ex. : `hasCuisine`) ayant pour domaine la valeur du tag dont il est dépendant (ex. : `v_restaurant`) et sa propre valeur pour rang (i.e. `v_seafood`). Cependant, seuls les tags qui possèdent plus de 100 occurences dans la base de données OSM, ainsi que ceux qui sont référencés dans le wiki OSM sont intégrés dans l'ontologie. Ce parti pris assure une meilleure qualité des tags considérés (ils sont *de facto* plus consensuels), mais écarte la grande diversité des contributions minoritaires significatives [12].

2.4. Game With A Purpose (GWAP)

Les auteurs de (Baglatzi Alkyoni *et al.*, 2012) adoptent une approche différente : ils proposent une méthode d'alignement des tags OSM sur l'ontologie de haut niveau DUL [13]. Le degré de similarité entre un tag OSM et un concept issu de l'ontologie DUL est mesuré via un jeu (*Game With A Purpose*, GWAP) : lorsqu'un contributeur ajoute un nouvel élément dans la base de données OSM, il ou elle se voit poser une série de questions à propos des caractéristiques dudit élément. Chacune de ces questions représente un concept de l'ontologie DUL. A titre d'exemple, la question "S'agit-il d'un objet physique tel qu'une rivière ou un stade ?" fait référence au concept `dul:PhysicalObject`. Cette technique dissimule la complexité de l'ontologie afin de faciliter l'utilisation de l'outil. Cependant, l'interprétation entre la ques-

9. http://dbpedia.org/
10. http://wiki.openstreetmap.org/wiki/OSMonto
11. http://do-roam.org/
12. *Exit* les erreurs d'orthographes et autres fautes de frappe
13. *DOLCE+DnS ULtralite*, http://www.ontologydesignpatterns.org/ont/dul/DUL.owl

tion et le concept interfère nécessairement entre la représentation du contributeur et le concept qui y correspond dans l'ontologie DUL. De plus, les utilisateurs ne sont pas sollicités dans le choix des concepts représentés dans l'ontologie de haut niveau.

2.5. OSM Semantic Network (OSN)

Enfin, le réseau sémantique OSN [14] (Ballatore *et al*., 2013) compte parmi les travaux les plus récents pour l'amélioration de la sémantique dans OSM. Les instigateurs de cette approche ont développé un outil qui explore les pages du wiki OSM (il existe, entre autres, des pages pour les tags et les clés) et les utilise pour générer un graphe dont les sommets sont les pages web et les arrêtes sont les hyperliens entrants et sortants. De surcroît, les concepts ainsi récupérés sont alignés sur les entités de Wikipedia et les concepts de l'ontologie LGD. Malgré tout, l'expertise humaine doit suppléer aux approximations de l'alignement automatique dans le but d'optimiser l'interconnexion des bases de connaissances. D'autre part, comme les auteurs l'ont souligné, des connexions doivent également être établies entre des bases de données géospatiales telles que GeoWordNet [15] ou Geonames [16] qui font aujourd'hui autorité. Ceci afin de tirer profit de toute l'information disponible sur les tags en consultation et d'éviter la redondance en modification.

2.6. Synthèse et motivations

Notre approche vise à assister la construction d'un modèle conceptuel pour OSM qui réponde à plusieurs problèmes. En premier lieu, afin de tenir compte de tous les éléments cartographiques (c'est-à-dire les représentations des objets physiques présents sur le territoire tels que les arbres, les aménagements urbains, le réseau routier, etc.) qui sont enregistrés dans la base de données, l'ontologie qui les modélise doit intégrer tous les tags existants qui les décrivent. La couverture des tags est donc un aspect de la plus haute importance pour prendre en considération la sémantique d'OSM de façon exhaustive. D'autre part, l'un des intérêts principaux de la création d'une ontologie est la réutilisation d'un vocabulaire commun dans l'optique de favoriser l'interopérabilité des systèmes informatiques. Par conséquent, l'interconnexion de l'ontologie OSM avec les autres bases de connaissances qui existent dans l'écosystème des données liées et ouvertes [17] est essentielle. Finalement, la base de connaissances OSM dont nous nous proposons d'appuyer la construction ne pourra être efficacement exploitée que si elle est solidement structurée. Dès lors, nous devons nous attacher à choisir une sémantique pertinente pour garantir l'expressivité des relations décrites dans le méta-modèle qui la sous-tend. D'autre part, bien conscients du problème de qualité soulevé par les données citoyennes, il nous semble indispensable d'instaurer

14. *OSM Semantic Network*, http://wiki.openstreetmap.org/wiki/OSM_Semantic_Network
15. http://datahub.io/dataset/geowordnet
16. http://www.geonames.org/
17. *Linked Open Data, LOD*, http://lod-cloud.net/

un système de gestion des versions de façon à s'assurer que chacune des modifications apportées à l'ontologie OSM puisse être annulée mais également pour servir de base à une analyse longitudinale des tags. Pour finir, dans la lignée des préceptes posés par le projet OSM, nous faisons le pari que la confiance en l'intelligence collective (O'Reilly, 2005) est cruciale pour encourager la responsabilité et l'autonomie des citoyens dans le processus de production de l'Information Géographique. Une attention toute particulière sera donc accordée à l'implication de l'utilisateur dans le méta-modèle que nous proposons.

Le tableau 1 livre une synthèse des cinq modèles de données pour les tags OSM étudiés en section 2, évalués à l'aune des cinq critères suivants : 1) exhaustivité de la couverture des tags, 2) interconnexion avec les bases de connaissances existantes, 3) implication de l'utilisateur, 4) expressivité et 5) gestion des versions. Plusieurs carences de la sémantique relative à l'information géographique participative (*Volunteer Geographic Information, VGI*) sont traitées par ces modèles. En effet, le GWAP est potentiellement capable d'intégrer tous les tags d'OSM. Le wiki OSM, quant à lui, implique de façon significative les utilisateurs dans le choix des tags et gère les différentes versions des pages du site web. Enfin, le réseau sémantique OSN est un premier pas vers une interconnexion substantielle entre les concepts issus de la folksonomie OSM et ceux qui proviennent d'autres bases de connaissances tout en offrant une taxonomie de tags relativement expressive grâce à des mesures de similarité entre ces concepts. Cependant, aucune de ces contributions ne répond à l'ensemble des cinq critères. C'est pourquoi nous proposons un modèle pour assister la construction collaborative d'une ontologie des tags OSM faite par et pour ses utilisateurs, qui tire le meilleur parti des travaux précédents et satisfasse l'intégralité des besoins évoqués plus haut.

Tableau 1. Comparaison des modèles de données pour les tags OSM

	Couverture	Interconnexion	Implication utilisateur	Expressivité	Gestion de Versions
OSM Wiki	Faible (2047/70$\overline{M}$ tags)	Faible (Hyperliens Wikipedia)	Forte (Cycle BRD)	Aucune	Oui (Type wiki)
LGD	Faible (2047/70$\overline{M}$ tags)	Moyenne (DBpedia)	Faible (Expertise descendante)	Faible (Subsomption clé-valeur)	Non
OSMonto	Faible (3000/70$\overline{M}$ tags)	Aucune	Faible (Expertise descendante)	Faible (Subsomption clé-valeur)	Non
GWAP	Forte (Potentiellement tous)	Faible (DUL)	Moyenne (Questionnaire)	Moyenne (DUL)	Non
OSN	Faible (2047/70$\overline{M}$ tags)	Forte (LGD, DBpedia, WordNet)	Faible (Ontologies descendantes)	Moyenne (LGD + Similarité entre tags)	Non

3. OF4OSM: un méta-modèle pour l'ontologie de tags OSM

Le modèle que nous présentons ici pour lier l'Ontologie et la Folksonomie d'OSM (OF4OSM) est composé de quatre parties qui correspondent aux quatre sous-sections suivantes : un modèle de représentation des tags (section 3.1), un modèle de relations entre tags (section 3.2), un modèle de similarité entre tags (section 3.3) et un modèle de révision de tags (section 3.4).

3.1. Une représentation des tags spécifique à OSM et conforme aux standards

La plupart des plates-formes d'UGC utilisent des tags pour classifier leur contenu. Dans (Lohmann *et al*., 2011), les auteurs font un état de l'art des ontologies de tags et proposent l'ontologie MUTO [18] qui vise à unifier les concepts fondamentaux sur lesquels repose l'activité de tagging et que l'on retrouve dans différentes ontologies qui font autorité telles que *Tag Ontology* [19], *Meaning Of A Tag (MOAT)* [20], *Common Tag* [21] ou *NiceTag* [22]. L'ontologie MUTO fait référence à des vocabulaires tels que SIOC [23] qui permet de lier un utilisateur qui associe un tag à une ressource (un *tagger*) à une communauté en ligne, mais également les schémas de métadonnées DCTERMS maintenus par l'initiative Dublin Core [24] ou encore le langage de représentation des connaissances SKOS [25] pour associer les concepts aux bases de connaissances respectant les standards du W3C [26]. Cependant, ce modèle ne correspond pas exactement à la structure des tags OSM (ils sont représentés par une paire clé-valeur) ni aux besoins évoqués dans la section précédente. Par conséquent, nous proposons un méta-modèle qui étend l'ontologie MUTO pour rapprocher l'ontologie et la folksonomie d'OSM : OF4OSM.

Dans l'ontologie OF4OSM, le concept `muto:Tag` est la super-classe du concept `of4osm:OSMTag`, comme décrit sur la figure 1. De cette manière, `of4osm:OSMTag` bénéficie des metadonnées héritées de `muto:Tag` : la description, la date de création et le créateur du tag. De plus, puisque `muto:Tag` est aussi un `skos:Concept`, `of4osm:OSMTag` est, par transitivité, décrit dans un langage de représentation des connaissances largement utilisé en Science de l'Information. Par ailleurs, le concept `of4osm:OSMTag` est associé à deux autres concepts représentant sa clé et sa valeur, respectivement `of4osm:OSMTagKey` et `of4osm:OSMTagValue`, par le biais de propriétés sous-classes de la propriété méréologique `dc:hasPart` :

18. *Modular Unified Tagging Ontology*, http://purl.org/muto/core
19. http://www.holygoat.co.uk/owl/redwood/0.1/tags/
20. http://moat-project.org/ns#
21. http://commontag.org/ns#
22. http://ns.inria.fr/nicetag/2010/09/09/voc.rdf
23. *Semantically-Interlinked Online Communities*, http://rdfs.org/sioc/ns#
24. *Dublin Core Metadata Initiative*, http://purl.org/dc/terms/
25. *Semantic Knowledge Organization System*, http://www.w3.org/2004/02/skos/core
26. *World Wide Web Consortium*, http://www.w3.org/

of4osm:hasOSMTagKey et of4osm:hasOSMTagValue. Cependant, la forme clé-valeur des tags OSM est équivoque. En effet, certaines clés sont faites pour être super-classes de leur valeur (ex. : amenity=school), tandis que d'autres sont faites pour représenter des propriétés booléennes (ex. : internet_access=no), selon les intentions du tagger (cf. section 2.2). Nous prenons le parti de laisser de côté cette ambiguïté. Au lieu de prendre chacune de ses composantes séparément, nous nous situons à un niveau d'abstraction supérieur pour considérer la paire clé-valeur comme un tout. Ainsi, le concept of4osm:OSMTag sera la brique élémentaire sujette à la classification dans l'ontologie de tags OSM.

Bien que la classe of4osm:OSMTag soit à la base de notre modèle de classification, nous avons besoin de la représentation de ses sous-parties, of4osm:OSMTag-Key et of4osm:OSMTagValue, qui peuvent être utiles pour interroger des services web d'information sur l'utilisation des tags. A titre d'exemple, TagInfo [27] est capable d'indiquer le nombre d'éléments cartographiques associés à un tag donné, ou bien s'il existe une page sur le wiki OSM qui correspond à ce tag. Ce type d'information peut s'avérer très utile au contributeur pour déterminer l'autorité d'un tag : s'il a un nombre d'occurences faible en base de données et pas de page dédiée sur le wiki, le contributeur devrait être plutôt dissuadé de l'ajouter dans l'ontologie. *A contrario*, si un tag est très largement utilisé et est documenté dans le guide de bonnes pratiques, il a très certainement sa place dans l'ontologie. Ces services donnent des informations relatives à un tag, mais également à une clé ou à une valeur. Prenons l'exemple suivant : un utilisateur souhaite ajouter le tag amenity=swimming_pool à l'ontologie. En interrogeant TagInfo, le système basé sur notre modèle peut signaler au contributeur que la clé la plus fréquemment associée à la valeur swimming_pool est leisure, qui est effectivement la clé qui fait autorité selon les directives OSM. Cette technique permet au contributeur de prendre des décisions avisées et, finalement, d'améliorer la qualité des données.

3.2. Une classification des tags plus expressive

L'un des principaux intérêts de la représentation de connaissances par le bias d'ontologies est l'expression des relations entre les concepts, la plus élémentaire d'entres elles étant la relation de subsomption. Malheureusement, les folksonomies sont intrinséquement plates : il n'existe pas *a priori* de relations hiérarchiques entre les tags. En conséquence, les différentes tentatives de sémantisation des tags OSM présentées en section 2 ont produit des taxonomies superficielles dont la profondeur est la résultante du seul rapport de subsomption entre clé et valeur. Pour parvenir à une hiérarchie plus profonde, nous proposons un concept abstrait (c'est-à-dire qui n'a pas d'instances dans la base de données OSM, qui ne peut pas être associé à un élément cartographique), of4osm:OSMAbstractTag, afin de servir de super-classe à d'autres tags (abstraits ou non) via la relation rdfs:subClassOf. Puisque of4osm:OSMTag

27. http://taginfo.openstreetmap.fr/

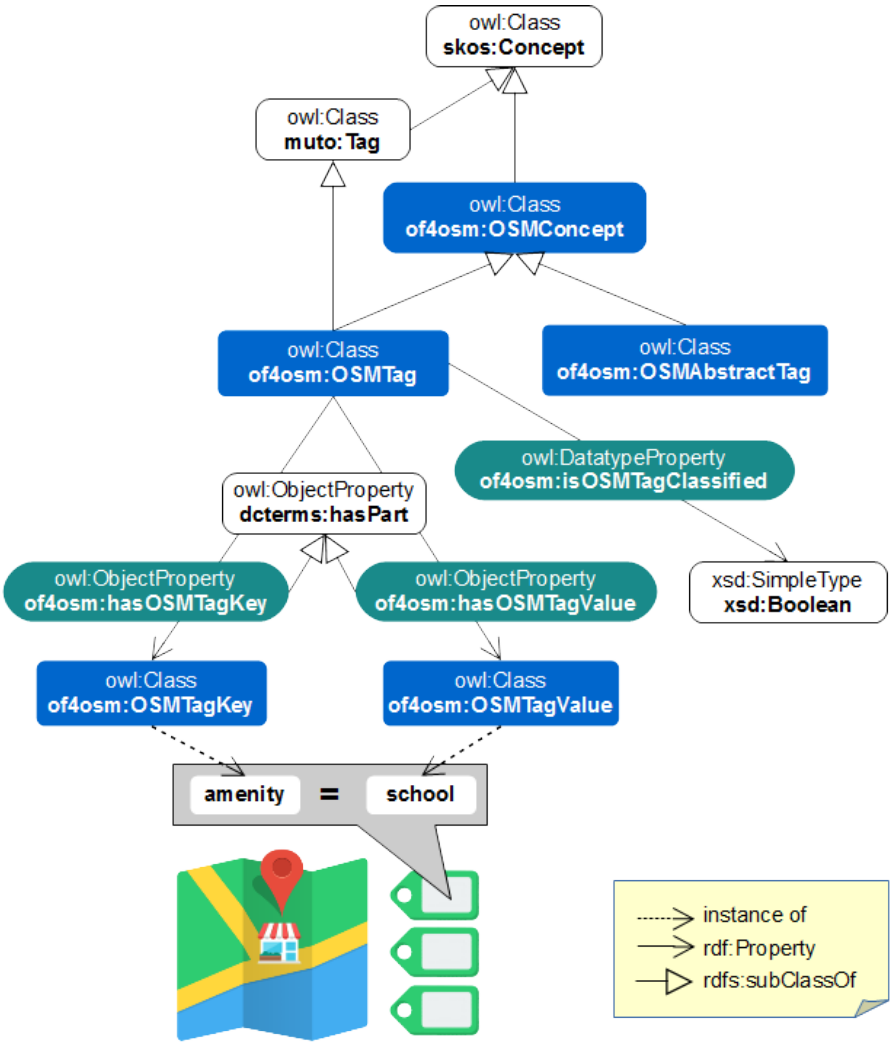

Figure 1. Modèle de relation entre tags autour du concept `of4osm:OSMTag`.

et `of4osm:OSMAbstractTag` sont les classes que les contributeurs vont agencer en une taxonomie, la capacité de subsomption qu'elles partagent est factorisée dans le concept parent `of4osm:OSMConcept`.

Pour illustrer l'intérêt de ce modèle, prenons l'exemple d'un contributeur OSM qui voudrait associer le tag simpliste `shop=bicycle` à un distributeur de chambres à air pour vélo. Si un système de suggestion de tags raisonne sur le réseau sémantique OSN dans lequel seules des relations de subsomption clé-valeur sont représentées (le concept `shop` est parent du concept `bicycle`), seuls les concepts qui ont également

pour parent le concept `shop` vont être retournés. Si on filtre ces résultats par nombre décroissant d'occurences dans la base de données OSM, on obtient les tags `shop=bakery`, `shop=clothes` et `shop=hairdresser`, lesquels présentent un intérêt pour le moins limité dans ce cas de figure. Avec une ontologie rudimentaire basée sur le méta-modèle OF4OSM, le concept abstrait `cycling_activities` pourrait être parent des tags `shop=bicycle`, `amenity=bicycle_repair_station` et `vending=bicycle_tube`, chacun d'eux étant documentés sur le wiki OSM mais ne partageant pas la même clé. Avec le même système de suggestion basé sur un algorithme de similarité de structure, cette nouvelle classification permet de retrouver les tags `amenity=bicycle_repair_station` et `vending=bicycle_tube`. L'utilisateur est cette fois à même de rafiner sa description du distributeur de chambres à air. Même si, à ce stade, notre méta-modèle ne décrit que des relations hiérarchiques entre les tags, il permet de générer une taxonomie des tags issus de la folksonomie OSM plus expressive que celles revues en section 2.

3.3. Compatibilité avec les mesures de similarité existantes et similarité subjective

Au-delà de la simple représentation des tags, il nous semble important de prendre en considération dans notre modèle une mesure de similarité qui provienne de la vision subjective du contributeur. En effet, nombre de mesures de similarité ont été développées, sur différents critères tels que l'interconnexion des pages du wiki OSM (cf. section 2.5), la lignée des tags (Mülligann *et al*., 2011), ou encore la similarité des instances géospatiales associées (Du *et al*., 2013). Pourtant, les premiers concernés, les membres de la communauté OSM, n'ont jamais eu l'opportunité d'exprimer un degré de similitude entre les tags. La figure 2 décrit le modèle de similarité entre tags que nous proposons et qui s'organise autour du concept central `of4osm:OSMTagSim` qui lie un tag à un autre selon une mesure de similarité spécifique représentée par la classe `of4osm:OSMTagSimType` et à laquelle un poids est associé via la propriété primitive `of4osm:hasOSMTagSimScore`. Cette représentation permet de tenir compte de différents types de mesures de similarité. Par exemple, la mesure de similarité subjective du contributeur OSM est repésentée par l'instance `of4osm:OSMSubjectiveTagSim`, issue de la classe `of4osm:OSMTagSimType`. De cette façon, la représentation du contributeur est intégrée au réseau de tags OSM ce qui répond aux éxigences de la politique participative du projet OSM.

Dans l'optique d'assister le contributeur lorsqu'il associe un tag à un élément cartographique, le modèle de similarité entre tags peut être exploité. Les auteurs de (Vandecasteele, Devillers, 2013) exposent une approche qu'ils ont implémentée dans un greffon, OSMantic [28], pour l'éditeur OSM JOSM [29]. Basé sur les mesures de similarité calculées dans OSN (cf. section 2.5), cet outil augmente le champ de texte dédié à l'ajout de tags de JOSM avec une liste de tags suggérés en fonction de leur degré de

28. http://wiki.openstreetmap.org/wiki/JOSM/Plugins/OSMantic
29. https://josm.openstreetmap.de/

similarité avec le tag entré par l'utilisateur. De plus, il alerte l'utilisateur qui voudrait associer un nouveau tag à une entité géographique qui en possède déjà d'autres qui présentent un degré de similarité très bas. Dans la même idée, nous fournissons un modèle générique pour encourager les contributeurs OSM à s'appuyer sur toutes les mesures de similarité disponibles car la diversité des points de vue est garante de choix avisés. Au-delà de l'aspect décisionnel, cela favorise la connexion entre l'ontologie de tags OSM et d'autres réseaux sémantiques.

Enfin, le processus de construction de l'ontologie de tags OSM doit pouvoir bénéficier de la comparaison entre, d'un côté, les nomenclatures officielles de caratéristiques spatiales du terrain et d'aménagements urbains et, de l'autre, les ontologies OSM. Cela permettrait de faire le pont entre les données gouvernementales et les données citoyennes qui présentent des caractéristiques très complémentaires comme souligné dans l'introduction de cet article. Techniquement, ce rapprochement peut être fait par le biais d'outils, les *matchers* d'ontologies, qui sont conçus pour déterminer des correspondences entre des concepts provenant de différentes ontologies. Les techniques d'alignement d'ontologies sont passées en revue de manière exhaustive dans (Euzenat, Shvaiko, 2007). Parmi elles, certaines techniques se basent sur la structure interne du modèle pour rechercher des similarités : l'expressivité de notre méta-modèle est ici déterminante pour permettre aux matchers de trouver des correspondences pertinentes comme montré en section 3.2. Ces alignements sont essentiels pour désenclaver les ontologies et favoriser leur interopérabilité. Pour illustrer ce point, nous avons aligné, à l'aide de matchers d'ontologies, d'un côté, une nomenclature officielle des aménagements urbains produites par l'INSEE [30] et, de l'autre, le réseau sémantique OSN. Entre autres, l'alignement qui en a résulté affiche une correspondence entre les tags `amenity=nursing_home` et `amenity=retirement_home`, issus d'OSM, et le terme `personne_agees_hebergement` de la nomenclature INSEE. Cette indication est précieuse pour le contributeur OSM : soit les termes `amenity=nursing_home` et `amenity=retirement_home` sont totalement redondants et il faut en supprimer un, soit ils partagent certaines caractéristiques communes qui doivent alors être factorisées dans une super-classe (abstraite avec `of4osm:OSMAbstractTag` ou non avec `of4osm:OSMTag`).

3.4. La gestion de versions comme support de la discussion entre contributeurs

Afin de conserver une trace des modifications effectuées sur l'ontologie OSM pour en améliorer sa qualité, nous proposons un modèle de revue des tags (figure 3) pour chaque tentative de classification via la classe `of4osm:OSMTagReview`. Cela permet d'enregistrer la date et l'identité du contributeur à l'origine de ce changement, ainsi que les relations (subsomption ou similarité) dans lesquelles le tag était impliqué avant la modification, de sorte que le système qui s'appuie sur notre modèle soit capable d'annuler toute action sur l'ontologie de tags. Dans le cas d'un profond désac-

30. *Institut National de la Statistique et des Etudes Economiques*, http://www.insee.fr/en/

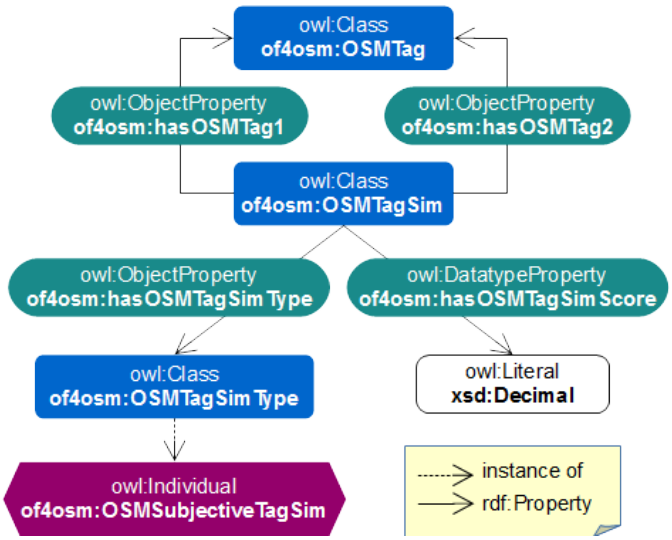

Figure 2. Modèle de similarité entre tags autour du concept `of4osm:OSMTagSIM`.

cord (de nombreuses annulations dans une période de temps courte [31]), l'utilisateur est incité à argumenter son choix par l'intermèdiaire d'une page de discussion sur le wiki OSM, soit pré-existante (dans ce cas, le système fournira l'url correspondante) ou non-existante (dans ce cas, le système se chargera de créer la page via l'API Wikipédia [32]), reprenant à son compte le cycle participatif BRD détaillé en section 2.1.

Au niveau des instances, lorsqu'un contributeur associe un nouveau tag à un élément cartographique, il est automatiquement ajouté et marqué comme inclassé au niveau des concepts dans l'ontologie de tags OSM via la propriété booléenne `of4osm:isOSMTagClassified`. Cette technique aide le contributeur à retrouver les tags qui n'ont pas encore été classés afin de limiter le nombre de tags non classifiés. On pourrait rétorquer que la simple absence d'une instance de la classe `of4osm:OSMTagReview` implique qu'un tag n'a jamais fait l'objet d'une classification. Ce n'est pourtant pas suffisant : si un utilisateur annule la première tentative de classification d'un tag sans en proposer de meilleure, le tag en question va retourner à sa position intitiale, sans instance de la classe `of4osm:OSMTagReview` associée. Dans ce cas de figure, le marqueur `of4osm:isOSMTagClassified=false` donne la possibilité à l'utilisateur d'indiquer à la communauté que le tag est revenu à son état antérieur mais qu'il a déjà été le sujet d'une ou plusieurs tentatives de classification (sa caractérisation est vraisemblablement délicate).

Par ailleurs, le modèle de révision de tag proposé ici est compatible avec la méthode de stratification de la confiance développée par les auteurs de (Exel, Dias, 2010)

31. Thresholds are to be determined in a future work.
32. http://www.mediawiki.org/wiki/API:Main_page

et peut être utilisé pour extraire des strates de données de plus ou moins bonne facture à partir de l'ontologie des tags OSM. Par exemple, en ignorant les tentatives de classification de faible qualité (des mesures de qualité des révisions seront développées dans de prochains travaux sur la base des mesures de qualité de données OSM mises au point par (Haklay *et al.*, 2010 ; Keß ler *et al.*, 2013 ; Rehrl *et al.*, 2013)), un environnement intégrant le méta-modèle OF4OSM est capable de générer des vues haute-qualité de l'ontologie OSM contenant uniquement les concepts clés. En ayant une vision globale dessinée par les tags centraux de l'ontologie sans pâtir du bruit des tags périphériques, le contributeur bénéficie d'un atout majeur pour éviter la redondance et l'incohérence dans le processus d'ingénierie des connaissances auquel il prend légitimement part.

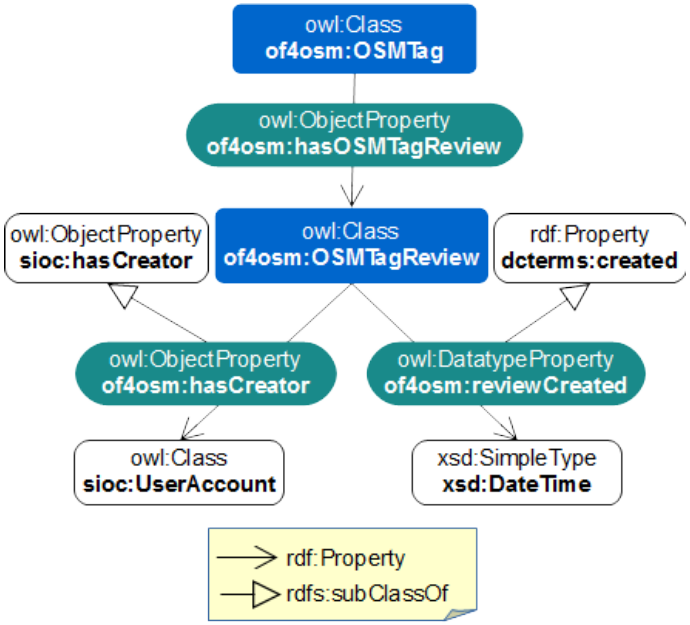

Figure 3. Modèle de revue de tag autour du concept `of4osm:OSMTagReview`.

4. Conclusion et perspectives

Nous avons introduit un méta-modèle qui vise à rapprocher la folksonomie de l'ontologie OSM. La méta-ontologie OF4OSM a pour but d'exploiter les tags OSM pour en faire une ontologie expressive faite par et pour les contributeurs OSM, qui préserve la flexibilité du système de tagging de contenu. Dans un souci de partage et de réutilisation de l'information, OF4OSM fournit un cadre de représentation des tags OSM qui s'inscrit dans l'écosystème des données ouvertes et liées en exploitant les vocabulaires standards du LOD. Cette représentation offre la possibilité aux contributeurs OSM de classifier les tags en une taxonomie dont la profondeur va bien au-delà de la simple subsomption clé-valeur qui prévalait jusqu'alors. Attentifs au problème de la qualité

des données qui va de pair avec une stratégie participative, notre modèle de similarité entre tags permet d'interconnecter les concepts issus de l'ontologie OSM et ceux issus d'autres bases de connaissances afin d'obtenir l'information la plus complète possible mais également pour prévenir la redondance. Enfin, notre modèle de revision de tags offre la possibilité d'enregistrer et d'argumenter toute nouvelle tentative de classification d'un tag dans l'ontologie, ceci afin d'encourager les contributeurs à s'engager de manière constructive dans l'élaboration d'une ontologie solide des tags OSM.

Les propositions exposées dans cet article sont les prémices de travaux pour l'élaboration d'un environnement dédié à la construction participative d'une ontologie des tags OSM. Cependant, beaucoup de travail reste à accomplir. Concernant la qualité des données, les travaux de M. Bishr et W. Kuhn sur les modèles de réputation des utilisateurs peuvent servir de point de départ à un système de contrôle de la qualité des contributions (Bishr, 2011). Du côté de l'interconnexion avec d'autres bases de connaissances, nous pensons que les mesures de similarité entre les tags du réseau sémantique OSN pourraient être améliorées en prenant en considération les différentes versions de ces pages web dans le temps (Ballatore *et al.*, 2013 ; Mülligann *et al.*, 2011). Pour finir, les modèles de comportement des contributeurs Wikipedia développés par D. Bégin sont une piste pour mieux appréhender les comportements des contributeurs OSM et, par là même, pour mieux comprendre les données produites par ces bénévoles (Bégin *et al.*, 2013).

Remerciements

Les auteurs remercient l'université Pierre-Mendès-France (Grenoble 2) et Minatec IDEAS Laboratory© pour le financement de la thèse d'Anthony Hombiat sur ce sujet.

Bibliographie

Auer S., Lehmann J., Hellmann S. (2009). LinkedGeoData: Adding a spatial dimension to the Web of data. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, vol. 5823 LNCS, p. 731–746.

Baglatzi Alkyoni, Kokla Margarita, Kavouras Marinos. (2012). Semantifying OpenStreetMap. In *The 11th international semantic web conference*, p. 39–48.

Ballatore A., Bertolotto M., Wilson D. C. (2013). Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems*, vol. 37, n° 1, p. 61–81.

Bégin D., Devillers R., Roche S. (2013). Assessing Volunteered Geographic Information (VGI) Quality Based On Contributors' Mapping Behaviours. *ISPRS*, vol. XL-2/W1, n° June, p. 149–154.

Bishr M. (2011). *Trust & reputation models for human sensor observations*. Thèse de doctorat non publiée.

Bruns A. (2006). Towards produsage: Futures for User-Led Content Production. In F. Sudweeks, H. Hrachovec, C. Ess (Eds.), *Cultural attitudes towards communication and technology 2006*, p. 275–284. Perth: Murdoch University.

Codescu M., Horsinka G., Kutz O., Mossakowski T., Rau R. (2011). Osmonto-an ontology of openstreetmap tags. *State of the map*. Consulté sur http://www.informatik.uni-bremen.de/~okutz/osmonto.pdf

Coleman D. J., Georgiadou Y., Labonte J., Observation E., Canada N. R. (2009). Volunteered Geographic Information : The Nature and Motivation of Produsers. *International Journal of Spatial Data Infrastructures Research*, vol. 4, p. 332–358.

Du H., Alechina N., Jackson M., Hart G. (2013). Matching Formal and informal Geospatial Ontologies. In D. Vandenbroucke, B. Bucher, J. Crompvoets (Eds.), *Geographic information science at the heart of europe*, p. 155–171. Cham, Springer International Publishing.

Euzenat J., Shvaiko P. (2007). *Ontology matching* (Springer-V éd.). Heidelberg (DE).

Exel M. V., Dias E. (2010). Towards A Methodology For Trust Stratification in VGI. , p. 2–5.

Golder S. a. (2006, avril). Usage patterns of collaborative tagging systems. *Journal of Information Science*, vol. 32, n° 2, p. 198–208.

Haklay M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, vol. 37, n° 4, p. 682–703.

Haklay M., Basiouka S., Antoniou V., Ather A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. , p. 1–13.

Ho S., Rajabifard A. (2010, mars). Learning from the crowd: The role of volunteered geographic information in realising a spatially enabled society. In *Gsdi conference proceedings.*

Keß ler C., Theodore R., Groot A. D. (2013). Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap. In *Geographic information science at the heart of europe*, p. 21–37.

Loenen B. V. (2012). Quest for a global standard for geo-data licenses. In *Spatially enabling government, industry and citizens: research and development perspectives*, p. 39 – 55. Needham, GSDI Association Press.

Lohmann S., Díaz P., Aedo I. (2011). MUTO. In *Proceedings of the 7th international conference on semantic systems - i-semantics '11*, p. 95–104.

McCall M. K., Minang P. A. (2005). *Assessing participatory GIS for community-based natural resource management: Claiming community forests in Cameroon* (vol. 171) n° 4.

Mülligann C., Janowicz K., Ye M., Lee W. C. (2011). Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In *Lecture notes in computer science*, vol. 6899 LNCS, p. 350–370.

Neis P., Zielstra D., Zipf A. (2011, décembre). *The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011* (vol. 4) n° 1.

O'Reilly T. (2005). *What Is Web 2.0 Design Patterns and Business Models for the Next Generation of Software.*

Rehrl K., Gröechenig S., Hochmair H., Leitinger S., Steinmann R., Wagner A. (2013). A Conceptual Model for Analyzing Contribution Patterns in the Context of VGI. In *Progress in location-based services*, p. 373–388.

Vandecasteele A., Devillers R. (2013). Improving Volunteered Geographic Data Quality Using Semantic Similarity Measurements. In *8th international symposium on spatial data quality*, vol. XL, p. 143–148.

Wal T. V. (2005). *Folksonomy Definition and Wikipedia.* Consulté sur http://www.vanderwal .net/random/entrysel.php?blog=1750

Weller K. (2007). Folksonomies and ontologies: two new players in indexing and knowledge representation. *Online Information 2007*, p. 108–115.