Vers une meilleure interopérabilité des données géographiques françaises sur le Web de données

Ghislain Auguste Atemezing, Raphaël Troncy

Vers une meilleure interopérabilité des données géographiques françaises sur le Web de données

Ghislain Auguste Atemezing, Raphaël Troncy

EURECOM, Sophia Antipolis, France
<auguste.atemezing@eurecom.fr>
<raphael.troncy@eurecom.fr>

Résumé : Les données géographiques sont cruciales pour une localisation précise et le geo-référencement de nombreux services qui sont disponibles sur notre planète. Au moment où le Web de données promeut l'idée d'avoir de plus en plus de données liées entre elles, il apparaît crucial de modéliser les données géographiques de manière efficace, en faisant usage de vocabulaires standards qui vont permettre de décrire de façon précise les formes géométriques. Nous observons que les initiatives existantes n'utilisent pas toujours la même approche pour la modélisation des objets géométriques. Dans cet article, nous commençons par passer en revue ces approches, qu'elles proviennent de la communauté *Linked Open Data* (LOD) ou de la communauté des Systèmes d'Information Géographiques (SIG). Nous cherchons à contribuer aux efforts actuels pour représenter les objets géographiques selon leur adresse, leur localisation, leurs propriétés fonctionnelles ou adminsitratives et leur formes géométriques. Nous proposons des alignements entre l'ontologie GeOnto et d'autres vocabulaires largements utilisés dans le Web de données (DBpedia, GeoNames, Schema.org, LinkedGeoData, Foursquare, etc.). Nous concluons par des recommandations quant à la publication de descriptions des objets géométriques sur le web de données.

Mots-clés : Ontologies, géographie, géométrie, Web de données, modélisation

1 Introduction

Le besoin d'orientation et de localisation va sans cesse croissant aidé en cela par des appareils mobiles de géo-localisation et des avancées dans le domaine des Systèmes d'Information Géographiques (SIG) donnant lieu à de nombreuses applications dans plusieurs domaines de la vie courante.

Des outils comme Google Maps [1] ou Google Earth [2] ont favorisé l'accès aux cartes numériques et tout l'éventail des services qui en découlent (visualisation d'un itinéraire, zoom sur une localisation pour avoir des Point d'intérêts (POI), etc.). Un autre phénomène qui a vu le jour ces dernières années est l'ouverture et la publication de données libres par des collectivités territoriales ou des états dans un objectif de transparence de la vie publique. Ce mouvement a débuté avec des pionniers tels que data.gov.uk (en Angleterre) ou data.gov (aux Etats-Unis) [3] suivis par de nombreuses villes et venant enrichir le nuage "Linked Open Data" (LOD) (Bizer *et al.*, 2009). De bonnes pratiques de publication des données sur le web ont été énoncées par Tim Berners Lee [4] : les objets ou **ressources** (pas seulement les documents) sont identifiés par des URIs, leurs descriptions sont interconnectées.

Pour la modélisation des données géographiques, il y a un certain nombre de questions qui se posent : (i) Qu'est ce qui dans le monde réel est un POI ? (ii) A quel détail de granularité se fait la représentation ? (iii) Quel est le but du modèle ? (iv) Quels éléments de modélisation de géométrie sont nécessaires pour représenter les données du monde réel ? Autant d'enjeux dont les réponses peuvent conduire à des choix de conception et de modélisation bien différents. Il va sans dire que sur le Web de données, le besoin d'interconnexion des données fait de l'interopérabilité des vocabulaires [5], et donc des sémantiques associées, une nécessité pour avoir de bons jeux de données et faciliter la consommation des données par les applications et les utilisateurs. Cet article a pour but de faire une revue de la manière dont les données géographiques sont représentées puis de proposer des alignements des vocabulaires utilisés dans le domaine géographique. L'Institut de l'Information Géographique et Forestière (IGN) a commencé à publier ses données selon les principes des données liées dans le cadre du projet Datalift [6]. C'est dans ce cadre que nous proposons quelques recommandations quant à la modélisation des objets géométriques complexes.

Ce papier est structuré de la manière suivante : la section 2 motive notre travail et propose deux scénarios utilisés par la suite. Dans la section 3,

1. `http://maps.google.com`
2. `http://earth.google.com`
3. `http://www.data.gov/catalog/geodata`
4. `http://www.w3.org/DesignIssues/LinkedData.html`
5. Nous utilisons le terme générique de "vocabulaire" pour désigner tant les ontologies que les taxonomies, thésaurus ou liste d'autorité.
6. `http://www.datalift.org`

nous présentons l'état de l'art sur la représentation des données géographiques et les schémas qui leur sont associés, à savoir ceux décrivant les entités et ceux décrivant la géométrie. La section 4 est consacrée à l'application des différents modes de modélisation sur nos scénarios. Nous présentons l'ontologie du domaine géographique français (GeOnto), ainsi que des alignements avec d'autres vocabulaires dans la section 5 avant de conclure et d'ouvrir quelques perspectives à ces travaux (section 6).

2 Motivation

L'engouement pour la géo-localisation fait des données géographiques une cible de prédilection tant dans le domaine traditionnel des SIG que dans le domaine du web de données. Deux grands facteurs ont contribué à la croissance des données géographiques sur le web : l'accès facile aux technologies de géo-codage grâce à la technologie GPS (Global Positioning Service) et le travail collaboratif des utilisateurs dans des projets comme Open Street Maps (OSM) ou GeoNames. Ces contributions collaboratives ont permis de rendre accessible sur une carte la position d'une ville, d'un point d'intérêt ou même d'un fleuve et les contributions ne cessent de croître avec cette possibilité d'accéder de façon gratuite aux données ainsi générées. De nos jours, avec l'arrivée des smartphones, plusieurs applications de micro-blogging consomment des données de localisation pour donner une plus-value dans leurs offres de service.

La communauté du Web de données s'efforce de publier des données de tout genre et de les interconnecter entre elles afin de créer le "nuage de données" (Cyganiak & Jentzsch, 2011), en prenant soin d'attacher des vocabulaires à chaque jeu de données, afin de faciliter l'interopérabilité et la compréhension pour les applications consommant ces données. Les données géographiques ne pouvant faire exception, on a vu émerger plusieurs vocabulaires différents pour décrire les objets géographiques. On a d'une part de grands volumes de données géographiques et de l'autre côté une diversité de vocabulaires et la complexité de la représentation de la géométrie.

La récente publication des statistiques sur l'état d'utilisation des vocabulaires dans le nuage de données[7], permet d'avoir une idée précise de la mise en place des bonnes pratiques telles que recommandées par Tim Berners-Lee. Dans l'état actuel, on dénombre 251 vocabulaires utilisés par 464 catalogues. Restreint au domaine géographique, les résultats

7. `http://stats.lod2.eu`

montrent que l'ontologie W3C Geo est la plus utilisée, suivie de l'ontologie `spatialrelations` de Ordnance Survey (OS). Le tableau 1 présente des statistiques de l'usage de classes et de propriétés des quatre ontologies les plus utilisées. Nous constatons que le prédicat `geo:geometry` est présent dans 1,322,302,221 triplets, juste après les prédicats `rdf:type` (6,251,467,091) et `rdfs:label` (1,586,115,316), soulignant bien l'importance des données géo-localisées sur le web.

Vocabulaires	**#Dataset**	**#Triplets**	**Accès SPARQL**
W3C Geo	21	15 543 105	Cache LOD
OS spatialrelations	10	9 412 167	OS dataset
GeoNames	5	8 272 905	Cache LOD
UK admin-geography	3	229 689	OS dataset

TABLE 1 – Quelques statistiques de l'usage de quatre vocabulaires géographiques dans le Web de données.

Tout au long de cet article, nous utilisons deux scénarios différents : modéliser les propriété et la géométrie de la *Tour Eiffel* et de l'arrondissement dans laquelle elle est située dans Paris, à savoir le *7^{e} arrondissement*. Le premier aura comme modèle de représentation un POINT et le second sera considéré un POLYGONE.

3 Etat de l'art

Dans cette section, nous tentons de répondre tour à tour aux trois questions suivantes : (i) Où sont les données géographiques ? (ii) Quels vocabulaires sont utilisés pour représenter les entités du monde réel ayant une quelconque adresse ou localisation ? (iii) Comment est modélisée la géométrie de ces objets qui sont ensuite affichés sur des cartes par des technologies de géocodage ?

3.1 Les sources de données

Les technologies de géo-codage permettent de faire comprendre aux machines que les noms des entités géographiques (ville, lieu, emplacement) font effectivement référence à un emplacement affichable sur une carte. Ainsi, les services de localisation utilisent un codeur géographique

(géo-codeur) pour transcrire le nom d'un lieu en un emplacement sur une carte. Ces géo-codeurs peuvent être commerciaux (e.g. Google maps) ou publics (e.g. OpenGeocoder [8]). Le web de données a pris avantage des technologies de géocodage existantes pour publier des millions de données géo-localisées. C'est ainsi que GeoNames [9] dispose de plus de 10 millions de données géographiques (soit 5,240,032 ressources, une ressource étant de la forme `http://sws.geonames.org/10000/`), DBpedia de 727,232 triplets ou encore LinkedGeoData de plus de 60 millions de triplets. Toutes ces données sont diverses de par leur méthode d'accès (point d'accès SPARQL, service web ou API), des entités qu'elles représentent ou des vocabulaires et taxonomies qu'elles utilisent.

Fournisseur	**#Triplets/Données**	**Accès aux données**
DBpedia	727 232 triplets	SPARQL endpoint
Geonames	5 240 032 ressources.	API
LinkedGeoData	60 356 364 triplets	SPARQL endpoint, Snorql
Foursquare	n/a	API
Freebase	8,5 MB	Service RDF Freebase
Ordnance Survey(Villes)	6 295 triplets	API Talis
GeoLinkedData.es	101 018 triplets	SPARQL endpoint
Google Places	n/a	Google API
GADM project data	682 605 triplets	Service Web
Projet NUTS Data	316 238 triplets	Service Web

TABLE 2 – Données géographiques et types d'accès sur le Web de données.

3.2 Les ontologies pour représenter des entités géographiques

Selon le type de modélisation rencontré, nous distinguons quatre grands groupes pour décrire des entités géographiques :

– (i) L'utilisation de codes de haut niveau (généralement en utilisant un ensemble fini) correspondant à des types précis. Par la suite, des sous-types sont attachés à ces derniers pour des objets appartenant aux types de haut-niveau. C'est par exemple l'approche suivi par

8. `http://www.opengeocoder.net`
9. `http://geonames.org/`

GeoNames [10] qui réutilise les concepts de SKOS [11] pour définir des codes et des classes (A, H, L, P, R, S, T, U, V). Chacune de ces classes correspond à un type précis. Par exemple, A correspond aux frontières administratives. Ainsi, les classes sont définies comme des `gn:featureClass a skos:ConceptScheme` et les codes sont des `gn:featureCode a skos:Concept`. L'avantage de cette approche est la réutilisation de skos, et donc l'interopérabilité des termes avec d'autres domaines. En revanche, cette manière de modéliser oblige à avoir des restrictions sur les propriétés ce qui rend l'alignement automatique plus hasardeux.

– (ii) La définition d'une ontologie propre sans faire référence à des vocabulaires existants. Les concepts sont structurés à l'aide de relations `rdfs:subClassOf`. Le vocabulaire de LinkedGeoData [12] utilise par exemple cette approche, dont les 1294 classes sont bâties autour d'un noyau de 16 classes de haut-niveau à savoir : `Aerialway`, `Aeroway`, `Amenity`, `Barrier`, `Boundary`, `Highway`,`Historic`, `Landuse`, `Leisure`, `ManMade`, `Natural`,`Place`, `Power`, `Route`, `Tourism` et `Waterway`. L'avantage, ici, est la prise en compte d'un grand nombre de POI (niveau détaillé des classes). En revanche, il n'y a aucune réutilisation des vocabulaires existants. L'ontologie GeOnto du projet ANR [13] du même nom suit également cette approche. La version "simplifiée" définit deux grands groupes (`EntiteTopographiqueArtificielle` et `EntiteTopographiqueNaturelle`) et contient 783 classes.
– (iii) : La définition de plusieurs ontologies par sous-domaine à représenter, avec la variante d'avoir une ontologie de haut-niveau servant de connexion entre les différents vocabulaires. L'un des avantages de cette approche est la réutilisation de ressources externes (ontologiques ou non-ontologiques), faisant ainsi un réseau interconnecté d'ontologies. C'est ainsi par exemple que Ordnance Survey [14] dispose d'un vocabulaire pour le découpage administratif [15] et un autre pour

10. `http://geonames.org/ontology/ontology_v3.0.rdf`
11. `http://www.w3.org/2009/08/skos-reference/skos.html`
12. `http://linkedgeodata.org/ontology`
13. `http://geonto.lri.fr/Livrables.html`
14. L'équivalent en Grande-Bretagne de l'Institut National Géographique de France
15. `http://data.ordnancesurvey.co.uk/ontology/admingeo/`

des besoins de statistiques [16]. Dans le même ordre d'idée, le projet GeoLinkedData [17] définit trois ontologies de domaine : le domaine des transports (aéroport, chemin, etc.) [18], le découpage administratif (province, communauté autonome, etc.) [19] et le domaine de l'hydrographie (types de cours d'eau, etc.) [20]. Cette approche réutilise largement les vocabulaires existants, avec comme principe un vocabulaire par sous-domaine. Cependant, l'importation entière des ontologies pour la réutilisation de peu de termes peut parfois être contraigante.

– (iv) : La définition d'une *liste à plat* ou faiblement structurée de catégories ou de point d'intérêts. C'est le cas des types de lieu utilisés par Foursquare [21] ou les catégories de Google Place [22]. Dans cette approche, il est nécessaire de convertir au préalable ces types en une taxonomie pour effectuer par la suite leur alignement avec d'autres vocabulaires.

3.3 Les modèles pour la géométrie

Les données géographiques pour représenter un emplacement ou un lieu sur le web sont modélisés de manières différentes (Salas & Harth, 2011). On distingue en particulier les formes suivantes :

– *le point* : forme classique de représentation d'un lieu par la latitude et la longitude dans un système géodésique donné (le plus utilisé sur le web étant le wgs84). Par exemple, GeoNames définit dans son modèle la classe `gn:Feature a skos:ConceptScheme` comme étant un `SpatialThing` du vocabulaire wgs84.

– *le rectangle*(“bounding box”) qui représente un emplacement par deux points ou quatre segments formant un rectangle geo-referencé. Dans ce type de représentation, le vocabulaire prévoit des prédicats supplémentaires pour les différents segments. Cette logique est présente dans l'ontologie de la FAO Geopolitical [23]. Un avantage est la simpli-

16. `http://statistics.data.gov.uk/def/administrative-geography`
17. `http://geo.linkeddata.es`
18. `http://geo.linkeddata.es/ontology/transportes.owl`
19. `http://geo.linkeddata.es/ontology/geopolitica.owl`
20. `http://geo.linkeddata.es/ontology/hydro-ontology.owl`
21. `http://aboutfoursquare.com/foursquare-categories/`
22. `https://developers.google.com/maps/documentation/places/supported_types`
23. `http://www.fao.org/countryprofiles/geoinfo/`

cité de la représentation (4 points) pour un polygone, au détriment de la représentation de géometries complexes.
- *les points dans une liste* : la forme géométrique est une région représentée par une collection de points, chacun décrit par un noeud RDF unique, qui à leur tour sont représentés en utilisant le vocabulaire wgs84 (lat/long). La classe `Node` est celle qui permet de connecter un point d'intérêt avec sa géométrie. Les points d'intérêts (POI) sont modélisés par des `Node` ou des surfaces (`waynodes`). Cette forme de représentation est utilisée par LinkedGeoData (Auer *et al.*, 2009).
- *les points* utilisant une seule propriété. L'objet est représenté par un groupe de ressources RDF appelées "courbes" (semblables au LineString de GML), qui est relié à l'objet décrit par la propriété *"FormeDe"*, et un attribut "ordre" pour préciser l'emplacement de chaque noeud dans la forme géométrique. Cette solution est celle appliquée dans GeoLinkedData (de León *et al.*, 2010). Ici, comme dans les *points dans les listes*, on prend en compte la géométrie complexe, mais avec comme inconvénient la génération de très nombreux noeuds anonymes.
- *les littéraux* : le vocabulaire prévoit un prédicat incluant une représentation GML de l'objet géométrique, qui est codé en RDF comme un littéral. C'est la solution adoptée par Ordance Survey (Goodwin *et al.*, 2008). L'avantage de cet approche est la compatibilité avec les standards de l'OGC. En revanche, le temps de réponse sera plus élevé pour répondre à des requêtes geospatiales.
- *la représentation structurée* de l'objet géométrique incluant aussi les objets complexes avec le vocabulaire NeoGeo [24]. Cette approche facilite le partage des points de la geométrie complexe et la séparation entre objet spatial et sa géométrie. L'inconvénient principal est la verbosité de la représentation.

4 Exemples de modélisation

Dans cette section, nous décrivons les différentes manières de représenter la *Tour Eiffel* et le *7e arrondissement* selon les vocabulaires décrits précédemment. Dans la suite, les préfixes utilisés sont définis par :

```
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos>.
```

geopolitical/resource/

24. http://geovocab.org/doc/neogeo/

```
@prefix lgdo: <http://linkedgeodata.org/ontology/>.
@prefix lgd: <http://linkedgeodata.org/triplify/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix
 virtrdf: <http://www.openlinksw.com/schemas/virtrdf#>.
@prefix dbpedia: <http://dbpedia.org/resource/>.
@prefix dbpedia-owl: <http://dbpedia.org/ontology/>.
@prefix dbpprop: <http://dbpedia.org/property/>.
@prefix fb: <http://rdf.freebase.com/ns/>.
@prefix gnr: <http://sws.geonames.org/>.
@prefix gn: <http://www.geonames.org/ontology#>.
@prefix geom: <http://geovocab.org/geometry#> .
@prefix spatial: <http://geovocab.org/spatial#> .
```

4.1 Tour Eiffel

Modélisation selon DBpedia et Freebase. DBpedia a la particularité d'avoir de nombreuses informations utiles sur cette ressource : architecte, jours d'ouverture, constructeur, etc). Nous observons qu'il y a 17 ressources différentes dans DBpedia faisant allusion à la Tour Eiffel à Paris.

```
dbpedia:Eiffel_Tower a dbpedia-owl:Building;
   a <http://schema.org/Place>;    (16 rdf:type au total)
   rdfs:label "Tour Eiffel"@fr;
   geo:lat "48.858299"^^xsd:float;
   geo:long "2.294500"^^xsd:float;
   geo:geometry "POINT(2.2945 48.8583)"
                          ; (geo:geometry n'existe pas)
   dbpprop:buildingType "Observation tower"@en;
   dbpprop:elevatorCount "9"^^xsd:int;
   dbpprop:location dbpedia:Paris;
   dbpprop:isoRegion "FR-75";
   dbpprop:architect dbpedia:Stephen_Sauvestre.
```

La relation `geo:geometry` est utilisée bien que celle-ci n'existe pas dans l'ontologie W3C wgs84. Selon Freebase [25], la même ressource est représentée de la manière suivante :

```
fb:en.eiffel_tower a fb:architecture.building;
   a fb:architecture.skyscraper;       (12 rdf:type in total)
   fb:architecture.skyscraper.height_with_antenna_spire_meters
      "324.0"^^xsd:float;
   fb:location.geocode [
```

25. http://rdf.freebase.com/

```
      fb:location.geocode.longitude "2.2946"^^xsd:float;
      fb:location.geocode.latitude "48.85839"^^xsd:float.
   ];
```

Modélisation selon GeoNames. Une modélisation classant la Tour Eiffel comme une structure commémorative au même titre que les statuts, définis par un point.

```
gnr:6254976 a gn:Feature;
   gn:name "Eiffel Tower";
   gn:alternateName "Eiffeltornet"@sv
                      ; (en 45 langues différentes)
   gn:featureClass gn:S [
      a skos:ConceptScheme;
      rdfs:comment "spot, building, farm, ..."@en.
   ] ;
   gn:featureCode gn:S.MMT [
      a skos:Concept;
      rdfs:comment "a commemorative structure or statue"@en.
   ];
   gn:countryCode "FR";
   geo:lat "48.8583";
   geo:long "2.29452".
```

Modélisation selon LinkedGeoData. La modélisation LGD est différente des précédentes car elle accorde une importance à la tour, et la modélise comme une séquence de 45 points utilisant pour cela une collection RDF (`rdf:Seq`) pour préciser l'ordre des points dans la liste.

```
lgd:way5013364
   a lgdo:Building, lgdo:ManMadeTower, lgdo:Attraction;
   rdfs:label "Torre Eiffel"@it; (13 langues différentes)
   lgdp:building:height "301";
   lgdp:importance "international";
   lgdo:hasNodes
     <http://linkedgeodata.org/triplify/way5013364/nodes>.
<http://linkedgeodata.org/triplify/way5013364/nodes> a rdf:Seq;
   rdf_1 lgd:node33388356;
   ...
   rdf:_10 lgd:node33388333;     (45 points du polygone)
```

4.2 7ème Arrondissement de Paris

Modélisation DBpedia. Dans cette modélisation, on remarque que les types `gml:_Feature` ou `grs:point` n'appartiennent pas à un vocabulaire OWL. De même, la propriéé `geo:geometry` n'existe pas dans le vocabulaire W3C wgs84.

```
dbpedia:7th_arrondissement_of_Paris a gml:_Feature;
      (gml n'existe pas en OWL)
   a <http://dbpedia.org/class/yago/1900SummerOlympicVenuEs>
   rdfs:label "7. arrondissementti (Pariisi)"@fi;
      (14 langues différentes)
   dbpprop:commune "Paris";
   dbpprop:département  dbpedia:Paris;
   dbpprop:région dbpedia:Île-de-France_(region);
   grs:point "48.85916666666667 2.312777777777778";
      (grs n'existe pas en OWL)
   geo:geometry "POINT(2.31278 48.8592)";
   geo:lat "48.859165"^^xsd:float;
   geo:long "2.312778"^^xsd:float.
```

Modélisation GeoNames. L'arrondissement est de 3ème ordre administratif et est représenté par un point, en plus des informations sur le nom alternatif, le code du pays et la population.

```
gnr:6618613 a gn:Feature ;
   gn:name "Paris 07";
   gn:alternateName "7ème arrondissement";
   gn:featureClass gn:A [
      a skos:ConceptScheme ;
      rdfs:comment "country, state, region ..."@en .
   ] ;
   gn:featureCode gn:A.ADM4 [
      a skos:Concept ;
      rdfs:comment
       "a subdivision of a third-order administrative division"@en .
   ];
   gn:countryCode "FR";
   gn:population "57410";
   geo:lat "48.8565";
   geo:long "2.321".
```

Modélisation LinkedGeoData. Le 7ème arrondissement est un quartier de la classe `lgdo:Suburb rdfs:subClassOf ldgo:Place`, mais dont la géométrie est représentée par un point, et non un objet géométrique complexe de type Polygone comme on pourrait s'y attendre.

```
lgd:node248177663
   a lgdo:Suburb;
   rdfs:label "7th Arrondissement"@en , "7e Arrondissement";
   lgdo:contributor lgd:user13442;
   lgdo:ref%3AINSEE 75107;
   lgdp:alt_name "VIIe Arrondissement";
```

```
georss:point "48.8570281 2.3201953";
geo:lat 48.8570281;
geo:long 2.3201953.
```

5 Alignement et discussion

Dans cette section, nous présentons la méthode utilisée pour l'alignement de l'ontologie GeOnto avec d'autres vocabulaires et les résultats obtenus. Nous terminons par quelques propositions pour une meilleure capture des connaissances sur la géométrie complexe dans le web des données.

5.1 Alignements

Une taxonomie existante du projet ANR appelée GeOnto, dont la version "simplifiée" disponible à l'IGN nous a permis de réaliser différents types d'alignement avec les vocabulaires existants. Cette taxonomie se caractérise principalement par les noms des identifiants en français, et la présence d'étiquettes en français et en anglais. Elle ne dispose pas d'annotations pour les classes. Cependant, la présence des labels anglais est très utile pour faciliter l'alignement de GeOnto avec le reste des vocabulaires qui disposent tous des mêmes caractéristiques. L'objectif étant de publier ladite taxonomie en précisant les équivalences entre les concepts au travers des vocabulaires existants pour une meilleure interopérabilité entre les ressources.

5.1.1 Méthodologie

Nous avons réalisé les alignements avec quatre vocabulaires au sens OWL (LGD, DBpedia, Schema.org et Geonames) et deux listes d'autorité (Foursquare, Google Place). Pour ces dernières, nous les avons automatiquement transformées en OWL en associant le type `owl:Class` à chacune des catégories. Pour chaque alignement effectué, nous nous limitons à considérer les relations du type `owl:equivalentClass`, et donnons les résultats fournies par l'outil Silk (Volz *et al.*, 2009). Nous utilisons principalement deux métriques de comparaison des chaînes de caractères : les distances de *Levenshtein* et *Jaro*. Ces métriques opèrent uniquement sur les labels des classes. Nous appliquons ensuite la moyenne des deux résultats obtenus.

Pour la génération du fichier final d'alignement des vocabulaires de taille réduite, nous procédons à une validation manuelle pour incorporer

Vocabulaire	**#Classes/Catégories**	**#Classes alignées**
LGD	`owl/Class` : 1294	178
DBpedia	`owl/Class` : 366	42
Schema.org	`owl/Class` : 296	52
GeoNames	`skos/Concept` : 699	287
Foursquare	`Catégories` : 359	46
Google Place	`Catégories` : 126	41

TABLE 3 – Métriques et nombre de classes effectivement alignées entre GeOnto et d'autres vocabulaires.

les relations de type `rdfs:subClassOf`. Le seuil retenu pour valider les résultats est de 100% pour les liens considérés corrects et supérieur à 40% pour les liens à vérifier. L'alignement avec Geonames a été spécial au sens que pour construire le fichier il a fallu procéder à diverses étapes : (i) utiliser Silk pour découvrir les codes `skos` pour les classes de GeOnto ; (ii) procéder à la vérification des liens sous le seuil de 70% ; (iii) charger le fichier généré précédemment dans un triple store ; (iv) et enfin, construire un nouveau graphe à l'aide de requêtes SPARQL de type `Construct` en appliquant les restrictions associées aux codes et classes de Geonames. Le tableau 3 présente les vocabulaires et taxonomies alignés avec GeOnto.

5.1.2 Evaluation/Résultats

En général, Silk donne de bons résultats avec des précisions au-delà de 80% (Google Place : 94%, LGD :98%, DBpedia :89%, Foursquare :92% et Geonames :87%.) Seul avec schema.org, nous avons obtenu une précision de 50% due à de nombreux "faux-positifs" comme par exemple `ign:Berge owl:equivalentClass schema:Park`.

Le tableau 4 donne un résumé des résultats obtenus pour l'alignement des différentes taxonomies avec le schéma de référence GeOnto.

L'alignement avec GeoNames est assez particulier car il nécessite de préciser les restrictions de la valeur du code d'appartenance de la classe de GeOnto. Ainsi, pour chaque classe de GeOnto, on génère le fichier GeoNames correspondant de la manière suivante :

```
geOnto:aGeoConcept a owl:Class;
  rdfs:label "a label"@en;
  owl:equivalentClass
    [ a owl:Restriction;
       owl:onProperty gn:featureCode;
```

	Précision	Recall	F-measure
Google Place	94%	100%	97%
LinkedGeoData	98%	93%	95%
DBpedia	89%	100%	94%
Schema.org	50%	100%	67%
Foursquare	92%	100%	96%
GeoNames	87%	100%	93%

TABLE 4 – L'évaluation de l'alignement des taxonomies en utilisant l'outil Silk.

```
      owl:hasValue gn:CODE ];
  rdfs:subClassOf  gn:Feature .
```

5.2 Recommandations sur les modèles géométriques

Dans les sections précédentes, nous avons vu les différentes formes de modélisation des objets géographiques. Pour la partie "symboligue" (feature), notre approche pour les données françaises est d'aligner GeOnto avec les vocabulaires existants, pour tirer profit au maximum des données déjà publiées sur le web de données. En ce qui concerne la représentation de la géométrie, nous souhaitons faire un choix en tenant compte des critères suivants :

- La prise en compte des objets géométriques complexes, en réduisant au maximum le nombre des noeuds anonymes dans les données pour faciliter la réutilisation des URIs.
- L'extension de GeOnto pour prendre en compte la géométrie, en réutilisant au maximum un vocabulaire existant.
- La fonctionnalité de prévoir des compatibilités pour les formats utilisés par les SIG, ainsi que la possibilité au besoin de partager des points pour certains scénarios.

Parmi les vocabulaires présentés jusqu'ici, NeoGeo semble être le vocabulaire qui pourrait répondre à nos critères sus-mentionnés. En effet, l'un des consensus des différents Vocamp sur la modélisation des objets géographiques est la séparation claire entre l'objet du monde réel représenté par une `Feature` et la géométrie associée. Cela se traduit par deux espaces de nom `spatial:Feature` et `geom:Geometry`, dont la relation `geom:geometry` sert naturellement de lien.

Le vocabulaire W3C wgs84 est ideal pour représenter les points. Cependant, pour ce qui est de la géométrie complexe (POLYGONE, MUL-

TIPOLYGONE, etc.), il n'est pas trop recommandable de représenter en RDF les points constituants ou membres de l'objet complexe (sauf peut-être pour des polygones de moins de 10 points). Il est plutôt souhaité de générer par des scripts à la volée le format pouvant intéresser les utilisateurs finals. Ces formats pouvant être GML, KMZ, KML, GeoJSON, SVG en plus de ceux usuels dans le web de données (Turtle, RDF/XML). L'avantage de cette approche est de permettre l'accès aux formats traditionnels du monde géographique tout en gardant la philosophie des bonnes pratiques de publication sur le web des données.

6 Conclusion

Nous avons présenté une étude des modèles existant sur le web de données pour représenter des données géographiques, ainsi que les modèles adoptés sur la représentation de la géométrie. Nous avons explicité les diverses formes de représentation des données géographiques sur deux scénarios (la tour Eiffel et le 7ème arrondissement de Paris) qui ont permis de mieux apprécier la diversité actuelle des représentations des données existantes sur le web de données. Par la suite, nous avons utilisé l'ontologie du domaine géographique (GeOnto) de la France dont son usage permettra la publication prochaine de larges contenu suivant les principes du LOD. Et pour faciliter son interopérabilité, nous avons procédé à des alignements (manuel et semi-automatique) avec différents vocabulaires et taxonomies. Enfin, nous avons préconisé une représentation plus structurée des objets géométriques complexes, avec le vocabulaire NeoGeo. De même nous préconisons aussi la négociation de contenue à la volée pour permettre de convertir les données dans les formats traditionnels des SIG.

Un travail futur serait une étude comparative entre les avantages des alignements effectués au niveau des schémas sur les alignements purement au niveau des données par les outils existants, sans usage préalable de tels alignements. Cette étude pourrait améliorer et permettre de développer des nouvelles métriques sur des outils de liaison de données intégrant aussi bien les similarités au niveau des concepts que d'autres relations propres du domaine de l'alignement d'ontologie. A court terme, nous espérons contribuer à faciliter la publication des données géographiques de la France en utilisant un schéma adéquat dans cette mouvance actuelle des données des gouvernements liées. Toutes les ressources mentionnées dans cet article sont disponibles à `http://semantics.eurecom.fr/datalift/ic2012/`.

Remerciements

Les recherches présentées dans cet article ont été partiellement financé par le projet ANR-2010-CORD-09 “Datalift”. Nous remercions OpenLink Software pour nous avoir facilité l’accès aux statistiques quant à l’usage de données géographiques sur le nuage LOD.

Références

AUER S., LEHMANN J. & HELLMANN S. (2009). LinkedGeoData - adding a spatial dimension to the web of data. In *Proc. of 8th International Semantic Web Conference (ISWC)*.

BIZER C., HEATH T. & BERNERS-LEE T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, **5**, 1–22.

CYGANIAK R. & JENTZSCH A. (2011). Linking open data cloud diagram.

DE LEÓN A., AL. V., VILCHES L. M., VILLAZÓN-TERRAZAS B., PRIYATNA F. & CORCHO O. (2010). Geographical linked data : a spanish use case. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS ’10, p. 36 :1–36 :3, New York, NY, USA : ACM.

GOODWIN J., DOLBEAR C. & HART G. (2008). Geographical linked data : The administrative geography of great britain on the semantic web. *Transactions in GIS*, **12**, 19–30.

SALAS J. & HARTH A. (2011). Finding spatial equivalences accross multiple RDF datasets. In *Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web*, p. 114–126, Bonn, Germany.

VOLZ J., BIZER C., GAEDKE M. & KOBILAROV G. (2009). Discovering and maintaining links on the web of data. In *International Semantic Web Conference (ISWC2009)*.