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, 2229 route des cretes, 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 tout ce qui peut être utile sur notre terre. A ce 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 des standards et vocabulaires existants afin de décrire de façon précise les formes géométriques. Jusqu'ici, les initiatives existantes n'utilisent pas toujours la même approche pour la modélisation des objets géométriques. Dans cet article, nous passons en revue les approches les plus utilisées tant dans les initiatives de Linked Open Data (LOD) que dans la communauté des Systèmes d'Information Géographique (SIG). L'intérêt étant de contribuer aux efforts actuels de représenter les objets géographiques qui possèdent des attributs liés à l'adresse, localisation ou emplacement. Nous proposons des alignements avec l'ontologie GeOnto, décrivant particulièrement la taxonomie dans le contexte français, et dont la publication prochaine selon les principes du Linked Data favorisera l'interopérabilité avec les vocabulaires existants (DBpedia, GeoNames, Schema.org, LinkedGeoData, Foursquare, etc.). Nous proposons enfin quelques recommandations des aspects à prendre en compte dans la publication 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 avec de nos jours les appareils de géo-localisation aidés en cela par les avancées dans le domaine des Systèmes d'Informations Géomatiques (SIG) et leurs applications dans plusieurs domaines de la vie. 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 la publication des données libres par les pays pour plus de transparence . Ce mouvement connu sous le nom du "Linked Open Data"(LOD) (Bizer *et al.*, 2009) a commencé avec des pionniers tels data.gov.uk (en Angleterre), data.gov (aux Etats-Unis) [3] et qui ne cesse de susciter l'engouement dans d'autres pays. Des bonnes pratiques de publication des données sur le web ont été énoncés par Tim Berners Lee [4], avec une place de choix sur leur modélisation pour être interconnectées et publiées, ainsi que la nécessité pour toute **"chose"** (pas seulement des documents) d'avoir un identificateur unique ou URI. Cet ensemble interconnecté des données constitue le web de données. Au même moment, plusieurs applications mobiles de géolocalisation sont mis à la disposition des utilisateurs dans des APPStore [5] faisant usage de la position actuelle de l'appareil pour faire des recommandations, des recherches basées sur la position voire des liens vers les comptes des réseaux sociaux. 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 [6], 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 étude de l'existant des données géographiques, de proposer des alignements du vocabulaire utilisé du domaine géographique en France [7],

1. `maps.google.com`
2. `earth.google.com`
3. `www.data.gov/catalog/geodata`
4. `http://www.w3.org/DesignIssues/LinkedData.html`
5. `http://fr.wikipedia.org/wiki/App_Store`
6. Nous utilisons le terme générique de "vocabulaire" pour désigner de façon générique tant les ontologies comme les taxonomies ou thésaurus, juste pour des besoins de simplification dans la rédaction.
7. L'Institut de l'Information Géographique et Forestière a commencé à publier ses données selon les principes du LOD, et c'est dans le cadre de cette initiative au sein du

et de proposer quelques recommandations sur la modélisation des objets géométriques complexes.

Ce papier est structuré comme suit : la section 2 présente la motivation de notre travail ainsi que les deux scénarios utilisés pour illustration. Dans la section 3, nous présentons l'état de l'art sur sur les données géographiques et les schémas qui leur sont associés, à savoir ceux décrivant l'entité et d'autres 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. Cette dernière s'achève par quelques recommandations, suivie d'une conclusion et des perspectives.

2 Motivation

L'engouement pour la géo-localisation fait des données géographiques une cible de prédilection tant dans son domaine traditionnel des Systèmes d'Informations (SIG) que dans le domaine du web de données, plus encore avec les initiatives actuelles de Linked Open Data (LOD). 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) et 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 de plusieurs endroits du monde ; 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. De l'autre côté, la communauté LOD s'efforce de publier des données de tout genre et de les interconnecter entre elles afin de créer le "nuage de données" (LOD cloud) (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

projet Datalift (www.datalift.org) que s'inscrit cette étude.

de données géographiques (ou "Big Data") et de l'autre côté, une diversité de vocabulaires et la complexité de la représentation de la géométrie.

2.1 Etude LOD Cloud

La récente publication des statistiques sur l'état d'utilisation des vocabulaires [8] sur le LOD [9], permet d'avoir une idée précise de mise en place des bonnes pratiques telles que recommandées par Tim Berners-Lee [10]. Pour ce qui concerne les vocabulaires et les données du domaine géographique, les résultats montrent que le vocabulaire W3C Geo est le plus utilisé, suivi de `spatialrelations` de Ordnance Survey (OS). La table 1 présente des statistiques dans l'usage des classes et propriétés de quatre ontologies les plus présent sur les jeux de données du Cloud. De plus, cette analyse signale que le prédicat `geo:geometry` est présent sur 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 géodonnées sur le web.

Vocabulaires	**#Dataset réutilisés**	**#Triplets sur le Cloud**	**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 admingeography	3	229 689	OS dataset

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

2.2 Scénarios

Tout au long de cet article, nous utilisons deux scénarios différents pour leur type de modélisation comme objet géométrique. Ainsi nous choisissons la très célèbre "Tour Eiffel" et l'arrondissement dans laquelle elle est située dans Paris, à savoir le 7^{eme} arrondissement. Le premier aura comme

8. Dans son état actuel, il y a 251 vocabulaires en usage pour 464 catalogues.
9. `http://stats.lod2.eu`
10. `http://www.w3.org/DesignIssues/LinkedData.html`

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 ? Et enfin (iii) comment est modélisée la géométrie de ces objets qui sont ensuite affichés sur des cartes par des technologies de géocodage ? Les réponses respectives à ces questions donnent lieu dans l'ordre aux sous-sections que nous détaillons par la suite.

3.1 Les sources de données

Les technologies de géo-codage permettent de faire comprendre aux machines que les noms des entités géographiques existantes (ville, lieu, emplacement, etc) 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 privés (exemple : Google maps) ou publics (exemple : OpenGeocoder [11]), tous permettant d'afficher les données sur des cartes, avec ou non des fonctions extra (comme celles de filtrage). Le web de données a pris avantage de ces technologies de géocodage existantes pour publier des millions de données. C'est ainsi que GeoNames [12] dispose de plus de 10 millions de données géographiques (soit 5 240 032 ressources [13], DBpedia de 750 millions de triplets ou encore LinkedGeoData avec près de 60 356 364 triplets, de même que d'autres initiatives LOD dans les pays (Grande-Bretagne, Espagne). Toutes ces données sont diverses de par leur accès (SPARQL endpoint, service web ou API), les entités qu'elles représentent ou les vocabulaires/taxonomies qu'elles décrivent.

11. `http://www.opengeocoder.net`
12. `http://geonames.org/`
13. Une ressource ici est de la forme `http://sws.geonames.org/10000/`

Fournisseur	#Triplets/Données	Accès aux données
DBpedia	727 232 triplets	SPARQL endpoint
Geonames	5 240 032 ressources.	API
LinkedGeoData Foursquare	60 356 364 triplets n/a	SPARQL endpoint, Snorql 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 modèles/ontologies pour l'entité géographique

Selon les types de modélisation rencontrés dans les données geospatiales, nous pouvons les regrouper en quatre grands groupes :

– (i) : Une des formes de structuration des entités est de définir des codes de haut niveau (généralement en utilisant un ensemble assez 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 dans l'ontologie de GeoNames [14], qui réutilise les concepts de SKOS [15] pour les codes et les classes (A, H, L, P, R, S, T, U, V), chacune des lettres correspondant à un type précis (exemple : A pour les 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 (ex : linguistiques), mais comme inconvénients on peut citer la restriction sur les propriétés et donc la difficulté pour un alignement automatique.

– (ii) : Une autre approche consiste à définir sa propre ontologie sans faire référence à des vocabulaires existants, et de

14. `http://geonames.org/ontology/ontology_v3.0.rdf`
15. `www.w3.org/2009/08/skos-reference/skos.html`

structurer aussi en grandes catégories autour desquelles les différentes autres classes sont des `rdfs:subClassOf`. Le vocabulaire de LinkedGeoData [16] utilise par exemple cette approche, dont les 1294 classes sont bâties autour d'un noyau de 16 classes de haut-niveau 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), mais par contre il n'y a pas réutilisation des vocabulaires existants.
Cette approche est un peu celle suivie par l'ontologie GeOnto du projet ANR [17], dont la version "simplifiée" définit deux grands groupes pour ses 783 classes. Elle distingue un premier groupe de haut-niveau composé de `EntiteTopographiqueArtificielle` et de `EntiteTopographiqueNaturelle`, comme leurs noms l'indiquent font reférence à des entités définies par l'homme et de celles se trouvant dans la nature respectivement. Le premier groupe étant subdivisé en 17 sous-catégories et le second en 4 autres sous-catégories.

– (iii) : D'autres approches différentes de celles décrites plus haut définissent un vocabulaire pour le 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 des ressources externes (ontologiques ou non-ontologiques), faisant ainsi un réseau interconnecté d'ontologies. C'est ainsi par exemple que Ordnance Survey [18] dispose d'un vocabulaire pour le découpage administratif [19] ; un autre pour des besoins de statistiques [20]. Dans ce même ordre d'idée, le projet GeoLinkedData [21] définit trois ontologies de domaine : le domaine des transports (aéroport, chemin, etc) [22] ; découpage administratif (pro-

16. `http://linkedgeodata.org/ontology`
17. `http://geonto.lri.fr/Livrables.html`
18. L'équivalent en Grande-Bretagne de l'Institut National Géographique de France
19. `http://data.ordnancesurvey.co.uk/ontology/admingeo/`
20. `http://statistics.data.gov.uk/def/administrative-geography`
21. `http://geo.linkeddata.es`
22. `http://geo.linkeddata.es/ontology/transportes.owl`

vince, communauté autonome, etc) [23] et le domaine de l'hydrographie(types de cours d'eau, etc) [24]. Cette approche réutilise intensément 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 déploré.

– (iv) : La dernière approche consiste à disposer d'une *liste à plat* des catégories des point d'intérêts. C'est le cas avec les types de lieu chez Foursquare [25] et les catégories de Google Place [26]. Pour cette approche, il est nécessaire de convertir au préalable ces types en une taxonomie pour 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 différentes manières (Salas & Harth, 2011). On distingue les différentes 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 W3C Geo [27]). Par exemple, GeoNames définit dans son modèle la classe `gn:Feature a skos:ConceptScheme` comme étant un "SpatialThing" du vocabulaire W3C Geo.

– *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 [28]. Un avantage est le nombre fixe nombre de points (4) pour représenter un polygone , au détriment de l'approximation de la géometrie complexe.

– *les points dans des listes* ; la forme géométrique est une région représentée par une collection des points, chacun de ces derniers décrit par un noeud RDF unique, qui à leur tour sont représentés en utili-

23. `http://geo.linkeddata.es/ontology/geopolitica.owl`

24. `http://geo.linkeddata.es/ontology/hydro-ontology.owl`

25. `http://aboutfoursquare.com/foursquare-categories/`

26. `https://developers.google.com/maps/documentation/places/supported_types`

27. `www.w3.org/2003/01/geo`

28. `http://www.fao.org/countryprofiles/geoinfo/geopolitical/resource/`

sant le vocabulaire W3C Geo (lat/long). La classe `Node` est celle qui permet de connecter un point d'intérêt d'avec la géométrie le représentant. Les points d'intérêts (POI) sont modélisés soit par des `Node` (c'est-à-dire points), soit par des surfaces (`waynodes`). Cette forme de représentation est suivie 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"* (FormadoPor dans l'ontologie originale de GeoLinkedData), 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 des "Blank Nodes".
– *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, mais par contre on un temps de réponse élevé pour les requêtes geospatiales.
– *Représentation plus structurée* de l'objet géométrique incluant aussi les objets complexes, approche du vocabulaire NeoGeo [29]. Cette approche facilite le partage des points de la geométrie complexe, et la séparation entre objet spatial et sa géométrie. L'inconvénient principal ici est la verbosité des données.

4 Exemples de modélisation

4.1 Tour Eiffel

Dans cette section, nous présentons les différentes manières dont la *Tour Eiffel* est modélisée dans les différents vocabulaires présentés dans les sections précédentes. Pour faciliter la lecture, nous donnons la liste des préfixes que nous utilisons et y faisons référence dans la suite de nos exemples.

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

29. `http://geovocab.org/doc/neogeo/`

```
@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#> .
```

4.1.1 Modélisation selon DBpedia

DBpedia a cette particularité d'avoir plus d'informations utiles sur la ressource (l'architecte, les jours d'ouverture, le constructeur, etc). Aussi, on se rend compte qu'il y a 17 ressources différentes dans DBpedia faisant allusion à la Tour Eiffel se trouvant 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 .
```

Il apparaît aussi l'usage de la relation `geometry` insérée par le triple store Virtuoso [30] mais qui n'est pas à l'origine disponible dans l'ontologie du W3C Geo. Selon Freebase [31], 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";
   fb:location.geocode [
      fb:location.geocode.longitude "2.2946"^^xsd:float ;
      fb:location.geocode.latitude "48.85839"^^xsd:float .
   ] ;
```

4.1.2 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)
```

30. `http://virtuoso.openlinksw.com/`
31. `http://rdf.freebase.com/`

```
   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" .
```

4.1.3 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 ("way" dans leur schéma), utilisant pour cela une collection RDF pour préciser l'importance de l'ordre des points sur la liste `rdf:Seq`[32].

```
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 ; (tous les 45 points ou polygone)
```

4.2 7ème Arrondissement de Paris

4.2.1 Modélisation DBpedia

Dans cette modélisation, on remarque que le type `gml:_Feature` n'est pas un vocabulaire au sens OWL[33] , de même que `grs:point`. Il apparaît également l'ajout de `geo:geometry` inexistante dans le vocabulaire W3C Geo.

```
  dbpedia:7th_arrondissement_of_Paris a gml:_Feature ;   (gml n'est pas OWL)
   a <http://dbpedia.org/class/yago/1900SummerOlympicVenuEs> (Classe YAGO)
   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'est pas OWL)
   geo:geometry "POINT(2.31278 48.8592)" ;
   geo:lat "48.859165"^^xsd:float;
   geo:long "2.312778"^^xsd:float.
```

32. http://www.w3.org/TR/rdf-schema/#ch_seq

33. http://www.w3.org/TR/owl2-overview/

4.2.2 Modélisation GeoNames

L'arrondissement est de 3eme ordre administratif et 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".
```

4.2.3 Modélisation LinkedGeoData

Paris 7ème 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 du type Polygone comme on se serait attendu pour cet objet.

```
lgd:node248177663
   a       lgdo:Suburb ;
   rdfs:label "7th Arrondissement"@en , "7e Arrondissement" ;
   lgdo:contributor lgd:user13442 ;
   <http://linkedgeodata.org/ontology/ref%3AINSEE>
           75107 ;
   lgdp:alt_name "VIIe Arrondissement" ;
   georss:point "48.8570281 2.3201953" ;
   geo:lat 48.8570281 ;
   geo:long 2.3201953 .
```

5 Discussion/Préconisation

Dans cette section, nous présentons la méthode utilisée pour l'alignement de GeOnto et les résultats issus de ce processus. Nous terminons la section par quelques propositions pour une meilleure capture des connaissances sur la géométrie complexe sur 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 des labels bilingues français-anglais. Elle ne dispose pas de commentaires 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 possibles 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 catégories (Foursquare, Google Place). Pour ces dernières, nous les avons automatiquement transformées en OWL en associant le type `OWL:Class` à chacune des catégories du fichier texte source. 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 *levenshteinDistance* et *jaro*. Ces métriques opèrent sur uniquement sur les labels en anglais des classes. Nous appliquons ensuite la moyenne des deux résultats obtenus des métriques précédentes.

Cependant pour la génération du fichier final d'alignement des vocabulaires de taille réduite, nous procédons à une validation manuelle pour incorporer les relations du 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, au regard du nombre pas trop important des données et des mesures de comparaison. L'alignement avec Geonames a été spécial en ce sens que pour construire le fichier il a fallu procéder en 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 des requêtes SPARQL de type `Construct` en appliquant les restrictions associées aux codes et classes de Geonames.

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/catégories effectivement alignées entre GeOnto et différents autres vocabulaires/taxonomies.

La table 3 présente les vocabulaires/taxonomies et les classes alignées 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 eu une précision de 50% due à de nombreux "faux-positifs". Comme par exemple `ign:Berge owl:equivalentClass schema:Park`.
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;
            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 au maximum des données déjà publiées sur le web des 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 "Blank Nodes" dans les données pour faciliter la réutilisation des URIs [34].
– La facile extension de GeOnto pour prendre en compte la géométrie, en réutilisant au maximum un vocabulaire existant.
– La fonctionnalité de prévoir des compatibilités pour formats en usage dans les SIG, ainsi que la possibilité au besoin de partager des points pour certains scénarios.

Des 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 [35] sur la modélisation des objets géographiques est la claire séparation entre l'objet du monde réel représenté par `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 Geo est ideal pour représenter les points. Cependant, pour ce qui est de la géométrie complexe (POLYGONE, MULTIPOLYGONE, 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 aux 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 les géo-données, 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 ; lesquelles 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

34. `http://www.ietf.org/rfc/rfc3986.txt`
35. `http://vocamp.org/wiki/Main_Page`

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.

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)*.