Comparaison de la nature d'objets géographiques

Nathalie Abadie, Ana-Maria Olteanu, Sébastien Mustière

Atelier OGHS'2007

Comparaison de la nature d'objets géographiques

Nathalie Abadie, Ana-Maria Olteanu et Sébastien Mustière

Laboratoire COGIT, Institut Géographique National
tel. : 33 1 43 98 80 00 fax : 33 1 43 98 85 81
{nathalie-f.abadie, ana-maria .olteanu,
sebastien.mustiere}@ign.fr

Résumé : L'appariement de données géographiques issues de bases de données hétérogènes consiste à mettre en correspondance les instances de classes des différentes bases qui représentent une même entité géographique du monde réel. Ainsi, parmi les instances de deux bases à mettre en correspondance, seules celles représentant des entités géographiques de natures similaires constituent de bonnes candidates à l'appariement. Nous proposons ici de comparer, à l'aide d'une mesure de similarité sémantique, la nature des objets géographiques à apparier pour détecter parmi tous ces objets, les meilleurs candidats à l'appariement. Afin de nous doter d'une mesure de similarité sémantique fiable et simple à mettre en œuvre, nous comparons ici trois méthodes nous permettant d'obtenir les valeurs de cette mesure.

Mots-clés : Mesure de similarité sémantique, appariement de données géographiques.

1 Introduction

Il arrive qu'un même espace géographique soit décrit par différentes bases de données géographiques, conçues et produites dans des buts et avec des moyens différents. C'est le cas notamment, à l'Institut Géographique National, de la BD TOPO ® et de la BD CARTO ®. Or, cette gestion indépendante des bases de données géographiques pose problèmes, tant pour le producteur que pour l'utilisateur des données. Elle implique, en effet, des efforts de saisie et de mise à jour répétés, et accroît de fait les risques d'incohérences entre les différentes bases (Sheeren, 2005). En outre, il est difficile d'effectuer des analyses utilisant conjointement des données de bases différentes. Pour pallier ces divers problèmes, une solution consiste à expliciter les relations existant entre ces différentes bases. Ce travail de mise en correspondance des bases de données géographiques comporte deux approches complémentaires, dans lesquelles la notion de proximité sémantique joue un rôle important : l'appariement des schémas des bases d'une part (Mustière *et al*., 2007), et celui des données elles-mêmes d'autre part (Olteanu *et al*., 2006).

L'appariement des schémas vise à déterminer quelles sont les classes de chacune des bases représentant un même phénomène du monde réel. L'appariement des données, quant à lui, consiste à identifier des liens de correspondances entre les

instances de classes des différentes bases qui représentent effectivement une même entité géographique du monde réel. Ainsi, l'appariement de schémas constitue une aide pour l'appariement des données, dans la mesure où il permet de restreindre le champ des recherches de correspondances entre instances de chacune des bases.

En l'absence d'identifiant universel pour les objets géographiques, l'appariement de données géographiques proprement dit se base généralement sur la géométrie des objets, ainsi que sur leurs toponymes. Ces critères sont suffisants lorsque l'on traite des classes dont les instances sont toutes de même nature. C'est le cas, par exemple, des classes « Cimetière » de la BD TOPO ® et de la BD CARTO ®. Cependant, il arrive que des choix de modélisation conduisent à représenter au sein d'une même classe un ensemble hétérogène d'entités géographiques. Les instances de cette classe sont alors spécialisées à l'aide d'un attribut « Type » ou « Nature ». Dans ce cas précis, les critères d'appariement énoncés précédemment s'avèrent parfois insuffisants. Ainsi, par exemple, deux instances de classes issues de bases différentes, et représentant une même entité du monde réel peuvent être géographiquement proches, mais avoir des dénominations différentes selon que l'on utilise leur toponyme d'usage ou leur toponyme officiel. C'est le cas de l'instance de la classe « Points Remarquables du Relief » de la BD CARTO ®, le « Col des joncs », qui a des coordonnées géographiques très proches de l'instance « Bizkartzu » de la classe « Oronyme » de la BD TOPO ®, mais un toponyme totalement différent. Une solution pour déterminer si ces deux instances représentent ou non une même entité du monde réel consiste à vérifier, en outre, si elles appartiennent à des catégories d'entités géographiques semblables ou pas. Ainsi, pour ces deux instances, des valeurs d'attribut « Nature » similaires, ou au contraire contradictoires, permettraient de lever toute ambiguïté et de décider si ces deux instances sont ou non de bons candidats à l'appariement. En l'occurrence, les attributs « Nature » des deux instances de classes citées ci-dessus ayant respectivement les valeurs « Col » et « Col, passage », ces deux instances représentent bien des entités géographiques de même catégorie, et doivent donc être appariées.

Or, la création de catégories d'entités géographiques relève d'un processus cognitif complexe (Mark, 1993). Ainsi, d'une base de données à l'autre, il arrive fréquemment qu'une même entité du monde réel soit répertoriée, au gré des différentes spécifications de saisie, dans des catégories différentes mais néanmoins proches d'un point de vue sémantique. Par exemple, le « Pic de l'Escarpu » est catalogué comme « Pic » au sein de la BD CARTO ®, et comme « Sommet » au sein de la BD TOPO ®. C'est pourquoi, dans le cadre du processus d'appariement, nous ne pouvons pas simplement comparer de façon systématique les valeurs des attributs « Type » ou « Nature » des instances de classes à traiter, mais nous devons évaluer le degré de proximité sémantique des différentes valeurs possibles de ces attributs. Pour ce faire, nous avons besoin d'une mesure de similarité sémantique. Nous comparons ici trois méthodes pour obtenir les valeurs de cette mesure.

2 Méthodes utilisées

Les tests effectués utilisent les instances des classes « Oronyme » de la BD TOPO ® et « Point Remarquable du Relief » de la BD CARTO ® de l'Institut Géographique National. Il s'agit de déterminer les objets candidats à l'appariement en comparant les valeurs de leur attribut « Nature », celui-ci pouvant prendre les valeurs « Cap », « Col », « Gorge », « Sommet », « Pic », etc., pour les objets de la classe « Oronyme », et « Cap, pointe », «Col, passage », «Défilé », « Pic, aiguille », etc., pour ceux de la classe «Point Remarquable du Relief ». Afin de définir une mesure fiable, nous avons testé trois méthodes pour mesurer la similarité sémantique, appliquées aux valeurs de cet attribut.

La première consiste à demander à des experts de fournir, pour chaque paire de concepts, une note, allant de 0 à 1, et traduisant leur perception de la proximité sémantique de ces concepts.

La seconde est basée sur une ontologie du domaine obtenue par extraction automatique à partir des textes des spécifications de saisie des bases de données géographiques de l'IGN (voir Fig. 1). Elle utilise la mesure de similarité de Wu-Palmer (Wu & Palmer, 1994), choisie pour sa simplicité de mise en oeuvre.

Enfin, la troisième est basée sur des résultats d'appariement d'objets géographiques: on admet qu'il existe une proximité sémantique entre les différentes valeurs possibles de l' attribut « Nature » des classes à traiter lorsque de nombreuses instances dotées de ces valeurs sont effectivement appariées grâce à leur géométrie ou à leurs toponymes.

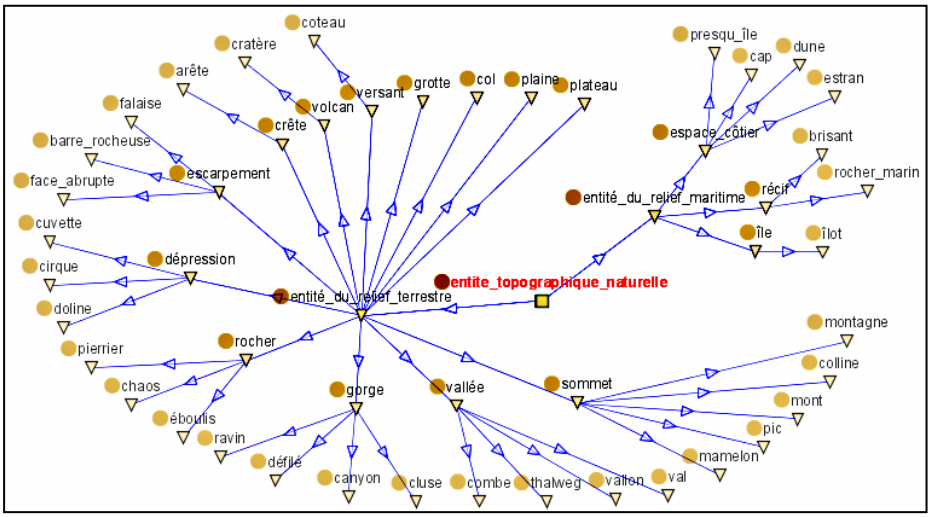

Fig. 1– Extrait de l'ontologie géographique utilisée (Visualisation avec Jambalaya)

3 Résultats

3.1 Difficultés méthodologiques

Une première difficulté méthodologique concerne les deux premières méthodes de calcul de similarité sémantique. En effet, elles utilisent une mesure qui s'appuie sur les valeurs de l'attribut « Nature » au sein des deux classes de bases de données à traiter, celui-ci pouvant regrouper plusieurs concepts géographiques à la fois. C'est le cas, notamment, pour la BD TOPO ® des valeurs « plaine ou plateau » ou « dune ou isthme », ou pour la BD CARTO ® des valeurs « cap, pointe, promontoire », « dune, plage», ou encore « plaine, plateau ». Ces regroupements de concepts géographiques ne correspondent pas toujours à des concepts plus génériques, mais obéissent à des nécessités d'acquisition ou de traitement des données. Cependant, ils compliquent considérablement les calculs de similarité sémantique, dans la mesure où il ne s'agit plus de déterminer une valeur de similarité sémantique pour une simple paire de concepts géographiques, mais pour une paire de groupes de concepts géographiques. Or, s'il semble évident que la paire de valeurs d'attributs « plaine ou plateau »-« plaine, plateau » devra logiquement prendre une valeur de similarité sémantique égale à 1, le cas de la paire « dune ou isthme »-« dune, plage » est déjà bien plus complexe à traiter. Ainsi, ayant été laissés libres dans leur notation, pour ce type d'exemple, les experts ont décidé des valeurs à attribuer de façon subjective. Dans le cas du calcul basé sur l'ontologie du domaine, le choix a été fait de développer le calcul et de conserver la valeur maximale obtenue. Ainsi, pour les différentes valeurs de similarité « dune »-« dune », «dune »-« plage », « isthme »-« dune », et « isthme »-« plage », on ne conservera finalement que la valeur de similarité « dune »-« dune », ceci afin de favoriser les tentatives d'appariement entre objets géographiques dont les attributs « Nature » possèdent au moins un terme commun.

Une autre difficulté est l'exploitation des valeurs de similarité proposées par les experts. En effet, aucune consigne de notation particulière n'ayant été donnée, chaque expert a fourni des notes très différentes de celles de ses homologues. De plus, ces notes présentent parfois l'inconvénient de ne pas respecter la propriété de symétrie d'une mesure de similarité. Enfin, si les évaluations de similarité sémantique entre paires de concepts restent cohérentes entre les trois méthodes de mesure, les notes moyennes fournies par les experts présentent des ordres de grandeur très différents des ceux obtenus avec les deux autres méthodes de mesure.

Par ailleurs, il est difficile de choisir, parmi toutes les mesures de similarité sémantique basées sur une ontologie du domaine, laquelle fournira les résultats les plus pertinents dans notre cas. Nous nous basons ici sur la mesure de Wu-Palmer, choisie en raison de sa simplicité d'implémentation. Celle-ci établit la valeur de similarité entre deux concepts grâce à la distance de leur plus petit généralisant. Cependant, elle présente une limite puisque l'on peut obtenir des valeurs de similarité de type $S(A, B) < S(A, C)$, B étant un concept fils de A, et C un concept frère de A (Zargayouna *et al*., 2004). Or, dans la mesure où l'on cherche à guider l'appariement d'objets géographiques, on ne peut, sous peine d'obtenir des résultats absurdes, favoriser des cas d'appariements entre types d'entités géographiques de natures incompatibles. Ainsi, lorsque le plus petit généralisant subsumant deux concepts dont on cherche à calculer la proximité sémantique est « entité topographique naturelle »,

cette mesure de similarité prend la valeur zéro. On s'assure ainsi d'éliminer d'éventuels appariements entre « entités du relief terrestre » et « entités du relief maritime », tout en conservant les propriétés d'une mesure de similarité.

Une dernière difficulté consiste à homogénéiser les valeurs de similarité sémantique obtenues avec ces trois méthodes de façon à établir des résultats facilement comparables. Ainsi, pour exploiter les résultats d'appariement, il convient de normaliser les chiffres obtenus, afin d'obtenir un ordre de grandeur comparable à ceux des chiffres obtenus via les deux méthodes précédentes. Cependant, il ne s'agit pas dans ce cas précis d'une mesure de similarité à proprement parler, mais d'une évaluation empirique des relations de proximité sémantique existant entre les catégories d'entités géographiques au sein desquelles sont regroupées les instances de classes des bases de données de l'IGN.

3.2 Intérêts comparés des sources de connaissances

3.2.1 Estimations des experts

Les notes attribuées par les experts (voir Table 1.) présentent l'avantage d'intégrer des connaissances de sens commun pertinentes mais non accessibles aux autres méthodes, comme la forme des entités géographiques, ou leur type de localisation. « En effet, l'ordinateur reste pour l'essentiel incapable […] d'une expérience sensible semblable à la nôtre des objets que sa connaissance réfère ». (Bersini, 2006) Les concepts de « dune » et de « colline » sont donc considérés par les experts comme proches d'un point de vue sémantique en raison de leurs aspects semblables, proximité sémantique qui ne traduisent pas les deux autres méthodes de mesure.

Ces notes présentent en revanche l'inconvénient de refléter des connaissances plus subjectives : une personne établit des connexions entre concepts géographiques en raison de sa relation à l'espace qui est avant tout une expérience personnelle. Ainsi, certains, du fait de leurs randonnées pédestres, tendent à rapprocher le concept de « cirque » de celui de « montagne », tandis que d'autres, ayant une expérience de géologues, considèrent qu'il s'agit d'une « dépression ». En outre, le concept de « rocher », dénué de tout contexte, pose également problème. En effet, en l'absence d'une connaissance fine des spécifications de saisie des bases de données géographique de l'IGN, de nombreuses confusions sont faites entre « rocher », entité du relief terrestre que l'on rencontre en zone de montagnes, et « récif », entité du relief maritime, que certains tendent à assimiler.

Table 1. Notes moyennes attribuées par les experts du domaine pour chaque paire de concepts géographiques

	Cap, pointe	Dune, plage	Ile	Récifs	Cirque	Col, passage	Cuvette, dépression	Pic	Sommet, crête, colline	Plaine, Plateau	Rochers	Vallée	Volcan
Cap	**1,00**	0,10	0,09	0,12	0,00	0,01	0,00	0,00	0,07	0,00	0,07	0,00	0,00
Dune, Isthme	0,17	**0,90**	0,11	0,08	0,00	0,02	0,00	0,02	0,09	0,00	0,01	0,00	0,00
Plage	0,10	**0,89**	0,17	0,13	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00
Ile	0,03	0,08	**1,00**	0,24	0,00	0,00	0,00	0,01	0,00	0,00	0,06	0,00	0,02
Récifs	0,17	0,18	0,21	**0,97**	0,00	0,00	0,00	0,00	0,00	0,00	0,17	0,00	0,00
Cirque	0,00	0,00	0,00	0,00	**1,00**	0,05	0,33	0,05	0,07	0,01	0,03	0,05	0,22
Col	0,00	0,00	0,00	0,00	0,02	**0,97**	0,02	0,06	0,04	0,03	0,02	0,01	0,01
Escarpement	0,05	0,00	0,02	0,12	0,15	0,05	0,06	0,19	0,16	0,01	0,33	0,03	0,11
Dépression	0,00	0,00	0,00	0,00	0,24	0,05	**1,00**	0,01	0,01	0,03	0,01	0,06	0,23
Pic	0,06	0,01	0,00	0,01	0,05	0,03	0,01	**1,00**	0,40	0,01	0,08	0,01	0,16
Sommet	0,02	0,03	0,00	0,01	0,06	0,03	0,01	0,43	**0,86**	0,04	0,06	0,01	0,15
Montagne	0,00	0,02	0,00	0,00	0,14	0,07	0,04	0,29	0,52	0,01	0,07	0,01	0,17
Crête	0,02	0,01	0,00	0,00	0,07	0,05	0,01	0,23	**0,82**	0,02	0,02	0,01	0,03
Plaine, plateau	0,00	0,00	0,01	0,00	0,03	0,02	0,02	0,01	0,01	**1,00**	0,01	0,11	0,02
Rochers	0,08	0,02	0,09	0,17	0,04	0,02	0,03	0,18	0,14	0,01	**0,90**	0,01	0,02
Vallée	0,00	0,00	0,00	0,00	0,08	0,13	0,13	0,01	0,01	0,18	0,01	**0,97**	0,02
Gorges	0,00	0,00	0,00	0,00	0,04	0,31	0,17	0,02	0,09	0,03	0,04	0,47	0,01
Volcan	0,00	0,00	0,04	0,00	0,15	0,01	0,14	0,16	0,16	0,01	0,06	0,01	**1,00**

3.2.2 Calcul à partir d'une ontologie du domaine

La méthode de mesure de similarité sémantique basée sur une ontologie présente l'avantage d'être simple à mettre en œuvre, dans la mesure où nous disposons déjà d'une ontologie du domaine (Mustière *et al.*, 2007). De plus, les modifications effectuées par rapport à la mesure de Wu-Palmer, nous garantissent de tout risque d'assimilation d' « entités du relief terrestre » avec des « entités du relief maritime ». Ainsi, la confusion, effectuée par les experts, entre « rocher » et « récif » n'est plus possible dans le cas de cette mesure (voir Table 2.). A l'inverse, au sein d'une même thématique, les résultats obtenus révèlent que cette mesure tend à rapprocher, bien plus que ne le font les experts, les concepts géographiques proposés. Or, dans la mesure où l'on cherche justement à détecter d'éventuelles proximités sémantiques entre concepts géographiques afin d'accroître le nombre d'objets candidats à l'appariement, ceci constitue un atout indéniable.

En revanche, il serait probablement plus pertinent de recourir à une mesure prenant également en compte, comme le font les experts, les propriétés des concepts géographiques, et en particulier leur forme. En effet, nous nous attachons ici à établir des correspondances sémantiques entre entités du relief dont la conformation

géométrique constitue la principale caractéristique (Mark & Gaurav, 2006). L'introduction de cette notion fondamentale en topographie permettrait d'obtenir des résultats plus contrastés selon que les concepts géographiques comparés possèdent ou non des formes semblables. Ainsi, les concepts de « plaine » et de « montagne » par exemple, dont la similarité est évaluée ici à 0,4 verraient cette valeur diminuée. En revanche, pour les concepts de « volcan » et de « montagne », qui avec une mesure de similarité de 0,4 sont considérés ici comme sémantiquement aussi proches l'un de l'autre que les deux exemples précédents, cette valeur augmenterait. Les concepts de « volcan » et de « montagne » caractérisés, contrairement au concept de « plaine », par leur forme proéminente, seraient donc considérés comme plus proches, et donc plus susceptibles de faire l'objet d'éventuels cas d'appariement.

Table 2. Valeurs de la mesure de similarité calculées à partir de l'ontologie de la Fig. 1

	Cap, pointe	Dune, plage	Ile	Récifs	Cirque	Col, passage	Cuvette, dépression	Pic	Sommet, crête, colline	Plaine, Plateau	Rochers	Vallée	Volcan
Cap	**1**	0,5	0,5	0,5	0	0	0	0	0	0	0	0	0
Dune, Isthme	0,5	**1**	0,5	0,5	0	0	0	0	0	0	0	0	0
Plage	0,4	**1**	0,4	0,4	0	0	0	0	0	0	0	0	0
Ile	0,5	0,5	**1**	0,5	0	0	0	0	0	0	0	0	0
Récifs	0,5	0,5	0,5	**1**	0	0	0	0	0	0	0	0	0
Cirque	0	0	0	0	**1**	0,34	0,34	0,57	0,67	0,34	0,34	0,3	0,34
Col	0	0	0	0	0,28	**1**	0,4	0,34	0,4	0,4	0,4	0,4	0,4
Escarpement	0	0	0	0	0,28	0,6	0,4	0,34	0,4	0,4	0,4	0,4	0,4
Dépression	0	0	0	0	0,28	0,5	**1**	0,4	0,5	0,5	0,5	0,5	0,5
Pic	0	0	0	0	0,57	0,6	0,4	**1**	**0,8**	0,4	0,4	0,4	0,4
Sommet	0	0	0	0	0,67	0,5	0,5	**0,8**	**1**	0,5	0,5	0,5	0,5
Montagne	0	0	0	0	**0,85**	0,6	0,4	0,67	**0,8**	0,4	0,4	0,4	0,4
Crête	0	0	0	0	0,57	0,6	0,4	0,67	**0,8**	0,4	0,4	0,4	0,4
Plaine, plateau	0	0	0	0	0,29	0,5	0,5	0,4	0,5	**1**	0,5	0,5	0,5
Rochers	0	0	0	0	0,34	0,5	0,5	0,4	0,5	0,5	**1**	0,5	0,5
Vallée	0	0	0	0	0,34	0,5	0,5	0,4	0,5	0,5	0,5	**1**	0,5
Gorges	0	0	0	0	0,28	0,6	0,4	0,34	0,4	0,4	0,4	**0,8**	0,4
Volcan	0	0	0	0	0,33	0,4	0,33	0,4	0,5	0,5	0,5	0,5	**1**

3.2.3 Résultats d'appariement

Cette méthode de mesure étant basée sur des résultats d'appariement, elle présente l'avantage de fournir des valeurs de similarité très pertinentes pour décider d'autres cas d'appariement. En effet, les chiffres (voir Table 3.), obtenus grâce à des appariements d'objets géographiques effectués sur des échantillons de bases de données géographiques de l'IGN, traduisent la connaissance très fine que les

opérateurs de saisie de ces bases ont des spécifications. Aussi cette méthode fournit-elle des valeurs de mesure de similarité reflétant très fidèlement une culture géographique commune. Or, les nouveaux échantillons de données que nous souhaitons traiter étant issus des mêmes bases de données géographiques, ils sont également imprégnés de cette culture. Ainsi, les concepts de « plaine » et de « montagne » se voyant ici attribuer une valeur de similarité nulle, toute tentative ultérieure d'appariement entre objets représentatifs de ces concepts s'avère inutile. A l'inverse, les paires de concepts considérés comme sémantiquement identiques dans les spécifications de saisie des bases de données géographiques obtiennent des valeurs très proches de 1, et feront à coup sûr l'objet de cas d'appariement. Enfin, des valeurs proches de 0,5, comme c'est le cas pour les paires de concepts « pic » - « sommet », ou « gorge » - « vallée », traduisent une ambiguïté très forte sur la sémantique de ces concepts géographiques au sein des différentes bases : un objet géographique catalogué comme « pic » dans la BD TOPO ® pourra voir son attribut « Nature » prendre indifféremment la valeur « pic » ou « sommet » dans la BD CARTO ®.

En revanche, pour garantir sa fiabilité, cette mesure de similarité doit être calculée à partir d'échantillons très larges. En effet, pour les valeurs de similarité proches de 0, comme c'est le cas ici entre « col » et « sommet », il reste difficile de savoir si l'on a affaire à un exemple d'appariement d'objets géographiques juste mais très marginal, ou à une erreur liée aux insuffisances de la géométrie et de la toponymie en matière d'appariement. Enfin, pour l'exemple proposé, de nombreuses paires de concepts n'ont pu être traitées, faute de données. En effet, l'échantillon de référence concerne l'ensemble du département des Pyrénées Atlantiques, choisi pour sa diversité de paysages. Néanmoins, on n'y rencontre ni « îles », ni « volcans », ce qui explique l'absence de valeurs pour ces concepts géographiques.

Table 3. Valeurs de similarité sémantique obtenues à partir de résultats d'appariement

	Cap, pointe	Dune, plage	Ile	Récifs	Cirque	Col, passage	Cuvette, dépression	Pic	Sommet, crête, colline	Plaine, Plateau	Rochers	Vallée	Volcan
Cap	**1,00**	0,00	X	X	0,00	0,00	X	0,00	0,00	0,00	0,00	0,00	X
Dune, Isthme	X	X	X	X	X	X	X	X	X	X	X	X	X
Plage	0,00	**1,00**	X	X	0,00	0,00	X	0,00	0,00	0,00	0,00	0,00	X
Ile	X	X	X	X	X	X	X	X	X	X	X	X	X
Récifs	X	X	X	X	X	X	X	X	X	X	X	X	X
Cirque	0,00	0,00	X	X	**1,00**	0,00	X	0,00	0,00	0,00	0,00	0,00	X
Col	0,00	0,00	X	X	0,00	**0,98**	X	0,00	**0,01**	0,00	0,00	0,00	X
Escarpement	0,00	0,00	X	X	0,00	0,00	X	0,00	0,00	0,00	0,00	0,00	X
Dépression	X	X	X	X	X	X	X	X	X	X	X	X	X
Pic	0,00	0,00	X	X	0,00	0,00	X	**0,58**	**0,06**	0,00	0,00	0,00	X
Sommet	0,00	0,00	X	X	0,00	**0,02**	X	**0,42**	**0,86**	0,00	0,00	0,00	X
Montagne	0,00	0,00	X	X	0,00	0,00	X	0,00	**0,03**	0,00	0,00	0,00	X
Crête	0,00	0,00	X	X	0,00	0,00	X	0,00	**0,04**	0,00	0,00	0,00	X
Plaine, plateau	0,00	0,00	X	X	0,00	0,00	X	0,00	0,00	**1,00**	0,00	0,00	X
Rochers	0,00	0,00	X	X	0,00	0,00	X	0,00	0,00	0,00	**1,00**	0,00	X
Vallée	0,00	0,00	X	X	0,00	0,00	X	0,00	0,00	0,00	0,00	**0,43**	X
Gorges	0,00	0,00	X	X	0,00	0,00	X	0,00	0,00	0,00	0,00	**0,57**	X
Volcan	X	X	X	X	X	X	X	X	X	X	X	X	X

4 Conclusion

Afin d'identifier les instances de classes de bases de données géographiques hétérogènes candidates à l'appariement, nous proposons de comparer, à l'aide d'une mesure de similarité sémantique, les valeurs des attributs « Nature » de ces instances. Pour ce faire, nous avons besoin d'une méthode de mesure de similarité sémantique à la fois fiable et simple à mettre en oeuvre. Nous avons donc comparé les valeurs de similarité sémantique obtenues à l'aide trois méthodes, pour un même échantillon de données. Une corrélation des différents résultats obtenus existe, qui tend à prouver que la méthode de mesure basée sur une ontologie, plus simple à implémenter, pourrait être mise en œuvre avec succès.

Ainsi, pour un même jeu de données, nous avons comparé des résultats d'appariement basés uniquement sur la géométrie et la toponymie, avec des résultats prenant en compte les valeurs de similarité sémantique de l'attribut « Nature » des objets géographiques à apparier, calculées à partir de l'ontologie du domaine. Dans ce dernier cas, les résultats obtenus ont une précision légèrement supérieure à celle obtenue via la première méthode. Le rappel, quant à lui, augmente de façon très

significative : d'environ 70% initialement, on obtient près de 100% grâce à l'ajout de ce troisième critère. Ces premiers résultats étant encourageants, nous envisageons d'étendre ces tests à des jeux de données plus larges ainsi que d'appliquer cette méthode à d'autres types d'objets géographiques tels les réseaux.

Références

BERSINI H. (2006) De l'intelligence humaine à l'intelligence artificielle. pp. 11.

MARK D.M. (1993) Toward a Theoretical Framework for Geographic Entity Types. In *Proceedings of the European Conference on Spatial Information Theory (COSIT'93)*. p. 270-283.

MARK D.M. & GAURAV S. (2006) Ontology of Landforms: Delimitation and Classification of Topographic Eminences. In *Proceedings of the International Conference on Geographic Information Science (GIScience'06)*. p. 129-132.

MUSTIERE S., ABADIE N. & LAURENS F. (2007) Appariement de schémas de BD géographiques à l'aide d'ontologies déduites des spécifications. In *EGC '07*., p. 22-27.

OLTEANU A.M., MUSTIERE S. & RUAS A. (2006) Matching Imperfect Data. In *Proceedings of the International Symposium on Spatial Data Accuracy Assessment in Natural Resources and Environmental Science (Accuracy '06),* Lisbon, p. 694-704.

SHEEREN D. (2005). Méthodologie d'évaluation de la cohérence inter-représentations pour l'intégration de bases de données spatiales – Une approche combinant l'utilisation de métadonnées et l'apprentissage automatique. *Thèse de doctorat, Université Paris 6.*

WU Z. & PALMER M. (1994). Verb Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics.* p. 133-138.

ZARGAYOUNA H. & SALOTTI S. (2004). Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. In *Actes de la conference d'Ingénierie des Connaissances (IC'2004).* p. 249-260.