Alignement d'un thésaurus sur GeoNames : retour d'expérience

Majid Khayari, Gilles Banzet

Alignement d'un thésaurus sur GeoNames : retour d'expérience.

Abdelmajid Khayari et Gilles Banzet
Equipe Terminologie-TAL
Inist-CNRS
2, Allée du Parc de Brabois
54514 Vandoeuvre-lès-Nancy Cedex
abdelmajid.khayari@inist.fr
gilles.banzet@inist.fr

Alignement de ressources terminologiques. Pourquoi ?

La mise en correspondance des ressources terminologiques répond à différents besoins[1,2] allant du souhait d'améliorer la recherche documentaire par l'interopérabilité sémantique qui permet à l'utilisateur d'accéder facilement à un même contenu à partir de plusieurs terminologies différentes (en termes de couverture linguistique, de granularité, etc.), à la volonté de gagner en visibilité en établissant des liens avec des ressources dont la notoriété n'est plus à démontrer. Le désir d'enrichir sa propre ressource par des données dont la production serait fastidieuse (traductions dans d'autres langues, définitions, etc.) ou des données impossibles à produire (nomenclatures, coordonnées géographiques et d'autres) peut également motiver une telle démarche. On citera également l'intérêt que procure cette information additionnelle pour l'indexeur dans la mesure où celui-ci peut, en consultant les ressources avec lesquelles l'alignement a été fait, lever une ambigüité en cas d'homonymie ou obtenir une traduction dans une autre langue, par exemple.

Les alignements produits de manière totalement manuelle ou en faisant appel à des algorithmes d'alignement peuvent être représentés et typés de manière assez fine à l'aide du vocabulaire SKOS[3] ; ainsi, deux concepts peuvent-ils être en correspondance exacte, approchante, générique, spécifique ou associative.

L'ajout de données géographiques à des documents ou à des ressources terminologiques directement sous forme de métadonnées ou via des alignements permet d'y inclure l'information géospatiale sous une forme normalisée et interopérable si cette information est issue de référentiels connus et reconnus tels que GeoNames. Les alignements seront autant de points d'entrée pour atteindre d'autres données (arborescence plus fine, données démographiques, traductions dans différentes langues, etc.). Le travail rapporté dans cette communication concerne l'alignement de la branche « Lieux » du thésaurus PACTOLS[4] sur GeoNames.

Alignement de la branche « Lieux » du thésaurus PACTOLS sur GeoNames

L'équipe Terminologie-TAL de l'Inist-CNRS a été sollicitée en 2016 par l'équipe en charge du thésaurus PACTOLS pour réaliser un alignement de la branche « Lieux » de ce thésaurus sur

---

[1] Jean Delahousse (2009). Sur l'alignement et la mise en correspondance de terminologies.
[2] McCulloch, Emma and Shiri, Ali and Nicholson, Dennis (2005) Challenges and issues in terminology mapping: a digital library perspective. Electronic Library, 23 (6). pp. 671-677. ISSN 0264-0473, http://dx.doi.org/10.1108/02640470510635755
https://strathprints.strath.ac.uk/2323/6/strathprints002323.pdf
[3] http://www.sparna.fr/skos/SKOS-traduction-francais.html
[4] Thésaurus spécialisé en relation avec la Préhistoire et l'Antiquité destiné à l'indexation et à l'interrogation du Catalogue collectif du Groupement de Service FRANTIQ (GDS 3378).

GeoNames. Malgré le nombre conséquent d'entrées (près de 21 000), la tâche a été jugée, de prime abord, faisable dans un délai raisonnable car un nom de lieu est un nom de lieu contrairement aux autres micro-thésaurus (Oeuvres, Sujets, etc.) pour lesquels l'exercice de conceptualisation puis d'alignement est assurément plus fastidieux.

Le micro-thésaurus « Lieux » comportait à l'époque plus de 20 700 entrées, correspondant majoritairement à des localités structurées par niveau administratif (commune, département/province, région, pays) ; s'y ajoutent quelques sites et/ou régions d'intérêt. Le micro-thésaurus ainsi que l'ensemble des PACTOLS sont consultables en ligne via OpenTheso[5], outil de consultation mais également d'édition ce thésaurus spécialisé en relation avec la Préhistoire et l'Antiquité, destiné dès son origine à l'indexation et à l'interrogation du Catalogue collectif du Groupement de Service FRANTIQ[6] (GDS 3378).

La base GeoNames[7] agrège quant à elle des données issues d'agences ou organismes nationaux (dont l'INSEE, data.gouv.fr et la SNCF pour la France) mais aussi de contributeurs individuels qui peuvent créer ou modifier le contenu de GeoNames[8]. La base comporte plusieurs millions d'entrées[9] correspondant à des localités, des toponymes, des oronymes, des hydronymes mais également des bâtiments publics des gares, divers sites, etc.

Chaque entrée a, entre autres,

- un identifiant unique GeoNames ;
- un *« name »* (nom préférentiel) ;
- éventuellement des *« alternateName »* (synonymes, traductions, translittérations, etc.) ;
- des coordonnées géographiques (latitude et longitude, élévation) selon le système de coordonnées WGS 84 (*World Geodetic System 1984*);
- une catégorie *« Feature Class »* représentée par une lettre de l'alphabet (GeoNames en possède 9 : « A » pour les niveaux administratifs, « P » pour les localités, « H » pour les cours d'eau et les lacs, « T » pour les montagnes, etc.) ;
- une catégorisation plus fine *« Feature Code »* de 2, 3 ou 4 lettres (GeoNames en possède 645, voir ce lien).
- un ou plusieurs codes pays ;
- un code pour les niveaux administratifs de rattachement administratif allant du plus générique (niveau 1) au plus fin (niveau 4).

La hiérarchisation des localités est assez fine allant de la commune au pays en passant par le département/la province et la région. Pour la France, l'arrondissement fait partie de la hiérarchie des localités.

Procédure d'alignement

Compte tenu de la taille de la base cible, l'alignement ne pouvait pas être réalisé par rapport à toute la base GeoNames mais en prenant les fichiers par pays disponibles au téléchargement sur le site de GeoNames. Ces fichiers sont au format txt tabulé et ont le double avantage d'être plus faciles à manipuler et de permettre de circonscrire les cas d'appariements non souhaités dus aux multiples homonymes.

[5] http://pactols.frantiq.fr/opentheso/
[6] Fédération et ressources sur l'Antiquité : https://opackoha.frantiq.fr/
[7] http://www.geonames.org/
[8] Ce type de contenu collaboratif dénommé « volunteered geographic information » par M.F. Goodchild n'est pas sans risques pour la qualité et la fiabilité des données. Voir cet article : https://www.sciencedirect.com/science/article/pii/S2211675312000097
[9] A la date du 25 mai 2019, la base affiche plus de 11,8 millions d'entrées.

L'utilisation d'outils d'alignement comme OnAGUI ou ITM-Match a été écartée car ces outils nécessitent des fichiers RDF ou OWL. Il existe bien un *« dump »* global en RDF mais qui a le double inconvénient d'être énorme (plus de 14 Go) et surtout pas à jour). La transformation des fichiers txt par pays en RDF a également été écartée car elle produit des fichiers trop volumineux pour pouvoir être importés facilement dans ces outils.

La procédure utilisée a consisté à transformer en XML[10] les fichiers txt par pays puis de mettre en place une chaîne de traitement (alignements, enrichissements, pré-validation, contrôles, etc.) utilisant des feuilles de style XSLT.

L'alignement a été réalisé en recherchant les termes d'un concept (préférentiel ou synonymes) du micro-thésaurus « Lieux » dans le champ *« name »* du fichier cible et uniquement dans ce champ afin de limiter les faux appariements dus à l'homonymie.

Pour chaque alignement, des données GeoNames sont récupérées (identifiant, rattachement administratif, coordonnées géographiques) ; le rattachement administratif permettant de faire une pré-validation des alignements (même nom, même rattachements administratif des deux côtés).

Un alignement ou des alignements GeoNames ?

Un premier alignement automatique a été effectué pour la branche France qui correspond, environ, à la moitié du micro-thésaurus « Lieux ». L'alignement a été envisagé, d'emblée, comme un alignement multiple car GeoNames propose plusieurs entrées pour un même lieu. Ainsi, toutes les localités villes et villages ont un équivalent ayant la catégorie « P » qui correspond au lieu proprement dit et un équivalent de catégorie « A » correspondant à son rattachement administratif : la commune. Dans la majorité des cas, c'est le même nom avec des catégories et un identifiant différents. Dans les autres cas, le nom de l'équivalent « A » est différent quand il y a eu regroupement de communes pour créer une nouvelle commune au sein de laquelle les anciennes communes deviennent des communes déléguées ou quand il s'agit d'un écart ou d'un hameau rattaché à une commune portant un autre nom.

Le choix de garder les deux alignements « P » et « A » a entraîné un travail de vérification des rattachements administratifs contenus dans GeoNames en prenant comme référence les pages Wikipédia. Ceci nous a permis de constater que la base GeoNames était moins réactive que Wikipédia. Cette dernière a créé très rapidement des pages pour les nouvelles régions, les nouvelles communes et mis à jour la majorité des pages correspondant aux anciennes communes ainsi que les noms des régions dans les pages des communes françaises. Des rattachements erronés ont également été trouvés dans GeoNames.

Alignements GeoNames, Wikipédia, DBpedia, Insee, RAMEAU, Wikidata et Loterre

En plus des alignements GeoNames majoritairement doubles « P » et « A », nous avons décidé d'inclure des liens Wikipédia_fr dans les entrées « Lieux » sous forme d'alignements car les pages Wikipédia que nous avions consultées à des fins de vérifications à la fois d'orthographe et de rattachements administratifs nous avaient semblé bien faites et fiables. Par extension, des liens vers les pages DBpedia ont également été générés. Ce choix a été conforté par le fait que GeoNames fait également référence à Wikipédia_en (en ligne) et, en plus, à DBpedia_en (et à l'Insee pour le fichier France) dans le *« dump »* RDF.

[10] Dans l'éditeur oXygen, le menu Fichier/Importer/Fichier texte permet de le faire de manière assez simple ; dans Excel, le menu Données/Fichier texte permet d'importer le fichier texte puis de le sauvegarder en XML mais le nombre de lignes est limité à 1048576.

Ultérieurement, d'autres alignements ont été produits en utilisant des requêtes SPARQL[11] ou l'API Wikidata via OpenRefine[12], il s'agit des alignements avec le Code Officiel Géographique (Insee), RAMEAU (BNF), Wikidata et Loterre qui expose une ressource contenant l'ensemble des communes françaises.

Corrections, enrichissements

Conjointement au travail d'alignement, un travail de correction du contenu du micro-thésaurus a été entrepris : modification de préférentiels obsolètes ou erronés, correction de rattachements administratifs erronés, suppression des doublons, remplacement des formes abrégées et/ou inversées par les formes en langage naturel, etc.

Un travail d'enrichissement a également été entrepris : ajout des traductions manquantes, des coordonnées géographiques GeoNames et de définitions extraites de Wikipédia. Initialement, des définitions ont été ajoutées manuellement dans les entrées « Lieux » qui ont été modifiées, notamment pour justifier nos interventions. Ultérieurement, l'ajout de définitions Wikipédia a été généralisé en utilisant DBpedia_fr. Cependant, le contenu de DBpedia n'étant pas à jour et n'étant pas le reflet du contenu Wikipédia, un travail de relecture des définitions[13] récupérées a été entrepris pour actualiser les définitions (notamment, le remplacement des anciennes régions par les nouvelles) et de mise en forme car « l'abstract » n'est pas toujours propre. Par la suite, c'est l'utilisation d'OpenRefine qui a été privilégiée car elle permet d'interroger le contenu actuel de Wikipédia et OpenRefine offre une interface de validation.

Bon à savoir

Voici quelques points techniques qui peuvent être utiles avant d'envisager un travail d'alignement avec GeoNames :

- Les données la base GeoNames sont actualisées en continu et mises à disposition sous forme de fichiers globaux ou par pays mais les données sont fournies sons forme de fichiers texte tabulé.
- Dans les fichiers .txt par pays, les entrées ont des identifiants GeoNames évidemment mais leurs niveaux administratifs de rattachement ne sont ni verbalisés ni donnés sous forme d'identifiants GeoNames mais sous forme d'identifiants propres au pays en question (identifiants Insee-La Poste pour la France, par exemple). Ainsi Strasbourg (P.PPLA) dont l'identifiant GeoNames est 2973783, est rattaché aux niveaux administratifs 44, 67, 678 et 67482 qui sont des codes Insee-La Poste. Le premier correspond à la nouvelle région Grand Est dont l'identifiant GeoNames est 11071622, le 2ème au département du Bas-Rhin dont l'identifiant GeoNames est 3034720, le troisième à l'arrondissement de Strasbourg-Ville dont l'identifiant GeoNames est 2973781, et le dernier à Strasbourg A.ADM4 dont l'identifiant GeoNames est 6441375. Ceci complique la tâche de reconstitution de l'arborescence. A noter qu'un fichier *« hierarchy.txt »* est disponible au téléchargement et présente l'arborescence des entrées de type « A » mais pas de type « P »…
- Une alternative intéressante aux fichiers txt consiste à extraire des données de FactForge qui inclut des données GeoNames et Wikipédia mais les données GeoNames ne sont pas forcément à jour et seules les pages anglaises de Wikipédia sont concernées.
- Compte tenu des homonymies très nombreuses, les « faux-positifs » peuvent être très nombreux également, d'où l'intérêt de croiser les informations (chercher dans GeoNames un lieu accompagné de son terme générique et ne garder que l'entrée GeoNames dont le

[11] https://query.wikidata.org/. Tutoriel : https://www.wikidata.org/wiki/Wikidata:SPARQL_tutorial/fr
[12] http://openrefine.org/
[13] A été considéré comme une définition, le début de la page Wikipédia extrait du DBpedia français en incluant le champ *« abstract »* dans la requête SPARQL

générique correspond…). Par ailleurs, le champ *« alternateName »* de GeoNames peut être très prolixe ; éviter, donc, de chercher des appariements en incluant les *« alternate names »* au risque de passer plus de temps à supprimer des faux appariements qu'à valider les bons.

- Le problème de l'apostrophe mérite également d'être signalé car nous avons eu plusieurs cas d'appariements partiels « P » uniquement sans les « A » correspondants car l'apostrophe était différente (par exemple, L'Arbresle pour « P » vs L'Arbresle pour « A »[14]) ou totalement absente. Il existe bien un champ « asciiname » pour remédier à ce problème d'apostrophe mais dans ce champ point de caractères accentués évidemment…
- A signaler également que GeoNames met parfois dans le champ *« name »* des variantes loco-régionales comme « el Voló » pour « Le Boulou » qui est mis en *« alternateName »…*
- Comme toute œuvre humaine, GeoNames n'est pas fiable à 100%...

Résultat final et implications

A l'issue des différents traitements et relectures, le fichier SKOS correspondant au nouveau micro-thésaurus « Lieux » est reconstitué en tenant compte des différentes données ajoutées, modifiées ou supprimées. Si la prise en compte des alignements et des enrichissements (définitions, traductions, synonymes, etc.) ne pose pas de problème particulier, il n'en va pas de même pour les corrections. En cas de correction d'un préférentiel erroné, l'ancien préférentiel a été mis dans un champ « terme caché » (*skos:hiddenLabel*) pour assurer les besoins d'interrogation de l'antériorité des références indexées avec ce terme. Si le préférentiel remplacé est une variante ou une forme ancienne, il a été versé dans un champ « synonyme ».

Concernant les relations hiérarchiques, celles-ci ont été entièrement recalculées compte tenu des corrections de génériques erronés, de l'introduction des nouvelles régions françaises (Hauts-de-France, Grand Est, etc.) et de la suppression de certaines anciennes régions (Alsace, Auvergne, Lorraine, etc.).

Les anciennes régions ainsi que le macro-découpage « France de l'Est », « France du Nord », etc. propre au micro-thésaurus n'ont pas été supprimés du thésaurus mais gardés et rattachés directement à « France métropolitaine » pour ne pas compromettre l'accès à l'antériorité de références indexées par ces termes. Dans GeoNames, les anciennes appellations (régions, communes, etc.) sont gardées et la lettre H est ajoutée à leur *« feature code »*.

Conclusion et perspectives

L'expérience relatée dans ce billet prouve, si besoin était, que tout travail terminologique est chronophage car les « moulinettes » facilitent le travail mais ne remplacent pas l'œil du rédacteur-relecteur seul à même de détecter des « faux-positifs » (nous avons eu de vrais alignements dus à des erreurs dans GeoNames…) ou d'autres anomalies, etc.

De plus, l'alignement sur GeoNames n'est jamais achevé car les données géographiques sont en constante évolution au gré des réorganisations territoriales qui impliquent une révision continue des alignements (de nouvelles régions ou communes sont créées et ne sont pas recensées immédiatement dans GeoNames puis le sont plus tard).

Enfin, le travail d'alignement exposé ici s'est doublé d'un travail d'enrichissement (ajout de définitions, de traductions, d'alignements avec d'autres ressources) et de correction de termes et/ou relations erronés ou obsolètes. Si les corrections qui concernent les relations hiérarchiques sont justifiées par la volonté de tenir compte de la modification des découpages territoriaux, elles

[14] Apostrophe dactylographique « ' » vs apostrophe typographique ou guillemet-apostrophe « ’ » qui est recommandé.

posent un problème de taille : celui de l'accès à l'antériorité des documents indexés par des entrées génériques (régions ou provinces) devant être remplacées par de nouvelles. Le modèle de données utilisé (SKOS) ne permettant pas de garder les entrées obsolètes ni de représenter leur renvoi sur les nouvelles entrées correspondantes, seule la possibilité de procéder à une ré-indexation du fonds documentaire pourrait rendre ce choix acceptable.