Sujet de thèse

Modèle de médiation pour la découverte et la réutilisation de contenus géographiques structurés

Laboratoire : LASTIG
Équipe : MEIG

Institut national de l'information géographique et forestière (IGN)

Discipline: Géomatique, Technologies des systèmes d'information
Mots-clés: Métadonnées, web sémantique, recherche d'information, mesure de similarité, catalogues, moteurs de recherche, données géographiques

Contexte

La découverte et la réutilisation de données géographiques prennent place dans un monde ouvert ; les utilisateurs, de différentes communautés, ne connaissent pas à l'avance la variété de contenus existants renvoyant à différents points de vue sur un espace géographique et à différentes techniques d'observation ou de modélisation (formats, nomenclatures, modèles conceptuels, résolutions). Un citoyen ou encore un rédacteur dans une administration étudiant un cours d'eau français devrait pouvoir explorer, choisir et réutiliser simplement des données les plus proches de son point de vue parmi celles disponibles : données d'autorité provenant de l'IGN, du BRGM ou de l'INSEE, données de DBpedia, données d'OSM, pour ne citer que ces sources. Un chercheur étudiant le climat urbain a lui besoin de sélectionner des données décrivant la topographie fine de la ville (comme la forme et les matériaux des bâtiments, la végétation, les altitudes, la perméabilité des sols), ainsi que des éléments sur les activités de l'homme (comme les sources de chaleur ou de pollution) à plusieurs dates données et sur des villes correspondant à plusieurs cas de figure. Une agence de services de paiement doit savoir sur quelles données locales peuvent s'appuyer des demandes de subsides européens, ou encore l'agence européenne de l'environnement doit savoir sélectionner et réutiliser des données locales et nationales pour alimenter des tableaux de bord européens.

Cette problématique est à rapprocher plus globalement de la recherche d'information sur le Web, qui a été révolutionnée par les moteurs de recherche permettant l'expression d'une requête simple, puis par l'exploration de réponses classées selon leurs niveaux de pertinences (scoring). Les prototypes actuels de moteurs de recherche de jeux de données géographiques sont souvent des moteurs plein texte dans les bases de métadonnées. Les textes des requêtes restent très brefs et concernent généralement la granularité spatiale ou temporelle des données recherchées [4].

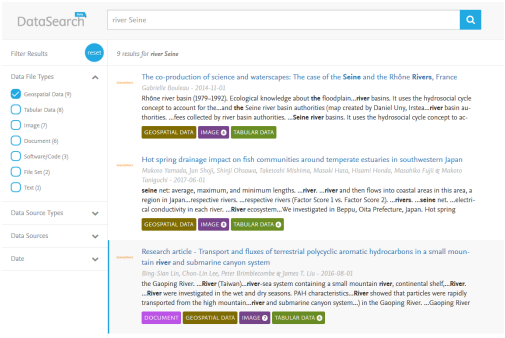

Figure 1. http://datasearch.elsevier.com

Malgré ces premiers prototypes, il demeure difficile pour un utilisateur de comparer diverses solutions proposées, d'estimer si les jeux de données trouvés répondent à son besoin sans avoir à les manipuler, de sortir d'un silo de ressources familières pour découvrir de nouvelles ressources et estimer s'il est capable de prendre en main les outils nécessaires à leur exploitation. Ce silo peut être lié par exemple à une technologie (portail Theia pour la télédétection), à un programme de financement (Copernicus), à un type de fournisseur (Géoportail) ou encore à une communauté (OSM).

L'IGN contribue au domaine général de la diffusion d'information géographique en ligne (données et services) depuis plus de 20 ans avec des travaux de recherche, d'expertise ou de développement et il opère depuis plus de dix ans le portail national français d'accès à l'information localisée, le Géoportail et participe au développement d'une infrastructure européenne de données géographiques INSPIRE [2]. À l'instar de plusieurs de ses homologues d'autres pays européens, l'IGN s'est vu récemment confier la mission de concevoir une plate-forme nationale d'intermédiation autour de l'information géographique. Une telle Géoplateforme vise à rassembler des communautés et des usages autour des données et des services géolocalisés.

Le laboratoire LaSTIG a développé par le passé avec les travaux de son ancienne équipe COGIT des compétences dans le domaine du web sémantique [1], [3], des moteurs de recherche [5], de la médiation de contenus structurés ou semi-structurés [9]. L'équipe MEIG dans laquelle s'effectuera la thèse reprend de nombreux thèmes du COGIT dans le domaine de la médiation, l'intégration et l'enrichissement de données géographiques fondées sur des représentations structurées de l'information géographique.

Sujet

Le sujet de cette thèse porte sur la capacité d'un moteur de recherche de données géographiques à traiter une requête initiale formulée à l'aide de termes du domaine de l'utilisateur ou de sens communs, à étendre ou raffiner cette requête, à identifier les réponses pertinentes possibles parmi les ressources et à les présenter à l'utilisateur de façon à lui permettre de les explorer suffisamment pour les comparer et choisir la plus satisfaisante. Les verrous abordés sont l'absence de modèle unifié permettant d'évaluer et comparer la valeur de différentes ressources en information géographique dans un contexte ainsi que le fossé d'expertise requis pour interpréter correctement les différentes métadonnées disponibles.

Les contributions visées de la thèse sont les suivantes :

- Une première contribution est un modèle multicritères de pertinence de contenus géographiques structurés qui utilise entre autres des distances de similarités entre les concepts de la requête et les concepts des ressources et qui prenne en compte des critères de qualité (cohérence, précision, exactitude, tant géométriques que sémantiques) ainsi que

d'expertise requise (comme le savoir-faire lié au paramétrage des ressources logicielles nécessaires à l'exploitation des sources).

- Une deuxième contribution est une restitution lisible par l'utilisateur des réponses identifiées par le moteur et de leur score de pertinence qui lui permette de comparer et choisir des ressources. Cette vue devra notamment croiser de façon fluide l'exploration des données et des métadonnées grâce aux techniques du web sémantique.

Concernant la méthode de travail, on ne visera pas une solution générique mais plutôt des solutions adaptées à des cas particuliers correspondant aux exemples décrits dans le contexte. Wikidata et les principales ontologies ou vocabulaires des cas d'utilisation considérés, y compris le vocabulaire INSPIRE, seront utilisés comme bases pivot pour l'expression ou la reformulation d'une requête et l'annotation des descriptions de ressources. L'approche envisagée pour la restitution consiste à charger des échantillons de données dans une représentation RDF qui permette d'avoir une représentation multiple et d'adapter la complexité à l'utilisateur, aussi bien en termes de structures géométriques manipulées (représentation de l'information) que de connaissance métier sous-jacente aux données. Les critères d'évaluation de la proposition seront 1/la capacité du modèle exploité par un utilisateur représentatif à obtenir automatiquement plus de réponses réellement pertinentes que les réponses obtenues par l'utilisateur seul face à l'ensemble des ressources, 2/à identifier les différentes ressources pertinentes identifiées par un expert, et à ne pas obtenir trop de réponses sans intérêt et 3/ enfin à permettre à l'utilisateur d'accéder effectivement à la ressource. Le retour de l'utilisateur sur la pertinence des résultats pourra être une source d'apprentissage tant sur le modèle de pertinence que sur celui de restitution. L'articulation avec les standards existants en géomatique et sur le Web ainsi que ceux qui émergent est essentielle pour bénéficier de la dynamique forte de ce domaine et sera faite par le biais d'échanges avec des développeurs et experts qui suivent ces standards.

Direction de thèse, encadrement de thèse

Directrice de thèse : Bénédicte Bucher, HDR (IGN)
Co-encadrement : Marie-Dominique Van Damme (IGN)

Collaboration : Sylvain Grellet (BRGM), Abdelfettah Feliachi (BRGM)

Contrat doctoral: Le contrat doctoral, d'une durée de trois ans, ouvre droit à une rémunération d'environ 1783 € brut (hors contribution aux frais de transports). Ce contrat peut comporter une charge supplémentaire d'enseignement, de diffusion de l'information scientifique et technique, de valorisation ou d'expertise.

Localisation: Equipe LaSTIG/MEIG, Institut national de l'information géographique et forestière (IGN), 73 avenue de Paris, 94 160 Saint-Mandé (région Parisienne, métro ligne 1 ou RER A).

Profil recherché: Géomatique, Systèmes d'information géographique, Web sémantique.

Candidature :

Le dossier de candidature est à transmettre avant le 31 mai par e-mail à benedicte (.) bucher @ ign (.) fr ainsi qu'à marie-dominique (.) van-damme @ ign (.) fr avec comme sujet « Candidature au sujet de thèse IGN médiation ». Il se compose des éléments suivants : votre CV, une lettre de motivation adaptée au sujet proposé, vos relevés de notes des deux dernières années d'études, l'avis du directeur de master (ou équivalent), un rapport de stage ou mémoire (facultatif), et le cas échéant des lettres de recommandations.

Les entretiens se tiendront entre le 3 et le 14 juin. La réponse sera fournie le 18 juin pour un démarrage de la thèse à la rentrée 2019.

Bibliographie:

[1] Atemezing, G., Abadie, N., Troncy, R., Bucher, B.. (2014) Publishing Reference Geodata on the Web: Opportunities and Challenges for IGN France. In (ISWC'14) 6th International Workshop on the Foundations, Technologies and Applications of the Geospatial Web (TerraCognita'14), Riva del Garda, Italy, October 19, 2014

[2] Bucher, B., Laurent, D., Janssen, P., (2019), Preserving semantics, tractability and evolution on a multi-scale geographic information infrastructure : cases for extending INSPIRE data specifications, Report of Eurogeographics-EuroSDR workshop on INSPIRE Data Extension, EuroSDR

[3] Hamdi, F., N. Abadie, B. Bucher and A. Feliachi (2014) GeomRDF: A Fine-Grained Structured Representation of Geometry in the Web, 1st International Workshop on Geospatial Linked Data (GeoLD 2014). In Conjunction with the 10th International Conference on Semantic Systems

[4] Kacprzak, E. , Koesten, L., Ibáñez, L.-D., Blount, T. Tennison, J., Simperl, E. (2018). Characterising Dataset Search – An Analysis of Search Logs and Data Requests. SSRN Electronic Journal. 10.2139/ssrn.3287149.

[5] Purves, R., Clough, P., Jones, C., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A., Vaid, S., Yang, B., (2007) The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet, in International Journal of Geographical Information Science, Volume 21, Number 7, page 717-745

[6] Smart P.D., Jones C.B., Twaroch F.A. (2010) Multi-source Toponym Data Integration and Mediation for a Meta-Gazetteer Service. In: Fabrikant S.I., Reichenbacher T., van Kreveld M., Schlieder C. (eds) Geographic Information Science. GIScience 2010. Lecture Notes in Computer Science, vol 6292. Springer, Berlin, Heidelberg

[7] Stuckenschmidt H., Visser U., Voegele T. J. Towards Intelligent Brokering of Geo-Information (2000). In Proceedings of the Urban Data Management Symposium

[8] Touya, G., Bucher, B., Falquet, G., Jaara, K., Steiniger, S., (2014) Modelling Geographic Relationships in Automated Environments, Abstracting Geographic Information in a Data Rich World, Burghardt, D. and Duchêne, C. and Mackaness, W. Ed., Lecture Notes in Geoinformation and Cartography, chap. 3, p. 53-82, Springer

[9] Van Damme M.-D., Olteanu-Raimond A.-M., Méneroux Y. (2019) Potential of crowdsourced data for integrating landmarks and routes for rescue in mountain areas. In International Journal of Cartography (accepté)