Création de bases de connaissances topographiques à partir de sources hétérogènes

Mots-clés: création et peuplement de bases de connaissances géoréférencées, intégration de données à références spatiales, extraction d'informations topographiques à partir de textes.

Contexte :

Les approches pour l'acquisition de données géographiques se sont pour la plupart attachées à produire des références spatiales directes (géométries vectorielles ou couvertures) soit, pour les plus récentes, par le biais de contributeurs bénévoles soit à l'aide de capteurs divers (GPS, téléphones, etc.). Peu de travaux, en revanche, se sont intéressés à l'extraction d'information géographique à partir de textes pour produire des données géoréférencées. Pourtant, certains textes offrent des descriptions du territoire et de ses infrastructures très précises et détaillées et constituent parfois la seule source d'informations disponible. En outre, nombre de documents règlementaires sur l'aménagement et les conditions d'usage des infrastructures publiques complètent ces descriptions avec des informations difficilement accessibles, hors traitement manuel des sources, via les modes d'acquisition de données géographiques classiques (règles de circulation, dates de mise en service, horaires d'ouverture, capacités d'accueil, etc.).

L'extraction d'information à partir de textes consiste à mettre en œuvre des techniques de traitement automatique du langage naturel (TALN) ou de fouille de textes (FdT) pour en extraire des informations structurées que l'on souhaite mettre à profit pour une application donnée (recherche d'information, analyses, visualisation, etc.). L'évolution récente des approches d'extraction d'information à partir de textes accompagne celle du Web de données qu'elles contribuent à alimenter en fournissant des outils pour la création automatique d'ontologies, le peuplement de bases de connaissances ou encore l'interrogation de bases de connaissances en langage naturel (Gangemi, 2013). A ces fins, elles s'appuient sur les possibilités offertes par les standards du Web de données et ses grandes bases de connaissances comme DBpedia ou Yago pour la désambiguïsation, la structuration et la vérification de la cohérence des informations extraites et la découverte de nouveaux faits par inférences (Suchanek, 2014).

Objectifs et questions de recherche :

Cette thèse vise à explorer les potentialités de l'extraction d'information à composante spatiale dans des textes pour construire et peupler une base de connaissances topographique à partir de sources hétérogènes décrivant un même territoire et selon différents points de vue. Les informations sur le territoire varient fortement d'un texte à l'autre, à la fois en termes de types d'entités géographiques décrites, de zone couverte, de temporalité, d'exhaustivité et de niveau de détail des descriptions fournies. Elles sont donc difficilement intégrables dans les modèles de données des SIG qui requièrent des données fortement structurées a priori. La convergence des possibilités offertes par l'extraction d'information et par les modèles de gestion de connaissances du Web sémantique fondés sur l'hypothèse du monde ouvert offre donc un cadre plus propice à la production et la gestion de connaissances sur le territoire à partir de textes (Chen et al., 2018).

L'objectif de cette thèse est donc méthodologique: il s'agit de proposer des approches opérationnelles pour permettre la construction, le peuplement et l'évaluation de la cohérence de bases de connaissances géoréférencées à partir de sources hétérogènes, notamment des textes, intégrant à la fois du référencement spatial direct et indirect pour le développement d'applications nécessitant de faire du raisonnement spatial selon ces deux modalités.

Il s'agira d'extraire, typer, désambiguïser et structurer les informations sur les entités topographiques décrites par des textes (noms, propriétés qualitatives, positions absolues et relatives dans l'espace, type d'entité géographique, etc.) pour les intégrer dans une base de connaissances et vérifier la cohérence des informations extraites, inférer de nouveaux faits, répondre à des requêtes sur le territoire, voire créer des sketch maps (Kim et al., 2016). Ceci suppose de proposer des solutions pour :

- Construire le plus automatiquement possible le vocabulaire associé;
- Adapter les approches existantes d'extraction d'information à composante spatiale à partir de textes à des corpus de divers types (anciens, techniques, réglementaires, etc.), caractérisés par différents types de structures et de vocabulaires selon le domaine;
- Représenter, stocker et manipuler des informations à composante spatiale qualitative (références spatiales indirectes, positionnement relatif, etc.) et à différents niveaux de détail selon les standards du Web de données ;
- Désambiguïser les entités topographiques extraites. Ceci suppose de :
    - Lier des entités topographiques éventuellement issues de plusieurs sources,
    - Prendre en compte des critères de liages dont la disponibilité pourra varier,
    - Prendre en compte les éventuelles variations des entités topographiques d'une source à l'autre (variations de nom, de propriétés, de temporalité, de niveau de détail, etc.),
- Détecter et corriger d'éventuelles incohérences spatiales ou temporelles dans les informations extraites, améliorer le typage des entités topographiques, inférer des relations spatio-temporelles entre entités topographiques, etc.

Approche:

L'approche adoptée pourra s'appuyer sur celle proposée dans (Keller et al., 2018) qu'il conviendra d'étendre afin d'améliorer chaque étape du processus de création de bases de connaissances. En particulier, les étapes réalisées manuellement dans cette première approche devront, autant que possible, être automatisés, notamment la construction de l'ontologie.

Une première étape sera donc de se doter d'un référentiel de base, de préférence doté de références spatiales directes et conforme aux bonnes pratiques du Web de données, afin de servir de support pour la désambiguïsation d'une partie des entités spatiales nommées extraites des textes. Celui-ci pourra être constitué à partir de données structurées existantes (bases de données vecteurs natives, gazetiers, cartes anciennes vectorisées, etc.) et devra être structuré à l'aide de vocabulaires adéquats pour la description des types d'entités représentées, de leurs propriétés, de leurs géométries, et de leurs relations afin de bénéficier des capacités de raisonnement associées aux standards du Web de données.

Une deuxième étape sera d'enrichir cette première base de connaissances par reconnaissance et résolution automatiques des entités topographiques mentionnées dans les textes. Pour les entités topographiques présentes dans le référentiel et dans les textes, ceci permettra de vérifier les niveaux de détail des descriptions disponibles et éventuellement enrichir la base de connaissances lorsque le texte y apporte un complément. Pour les entités topographiques non présentes dans la base mais

mentionnées dans les textes, ceci conduira à enrichir la base en ajoutant ces entités et leur description.

La réalisation de cet objectif nécessitera de proposer une chaîne de traitement opérationnelle et générique pour extraire des informations du type : entités topographiques, types d'entités topographiques, relations spatiales, propriétés et valeurs de propriétés. Celle-ci pourra s'appuyer sur une première proposition présentée dans (Lamotte et al., 2020). Elle supposera en outre d'enrichir les vocabulaires créés lors de la première étape afin de permettre la représentation de types d'entités topographiques ne figurant pas nécessairement dans la base, et celle de leurs propriétés spatiales, notamment les relations spatiales permettant leur localisation relative par rapport aux entités topographiques pour lesquelles on dispose de géométries.

Une fois extraites et typées, les entités topographiques devront être comparées à celles déjà présentes dans la base de connaissances afin de permettre de décider de l'opportunité de leur ajout ou de l'enrichissement de leur description dans la base. Cette étape de résolution des entités topographiques extraites du texte nécessitera de proposer des approches dédiées selon les types d'entités traités. Par exemple, certains types d'entités topographiques pourraient figurer dans les textes et dans la base de connaissances de référence, mais pas directement sous la forme d'objets, instances d'un concept géographique communément utilisé en langage naturel et devront être préalablement explicités pour faciliter leur résolution. En outre, à cette étape, il sera impératif de tenir compte de la temporalité des informations extraites dans le processus d'enrichissement de la base.

Une dernière étape du processus de construction de base de connaissances consistera à permettre le contrôle de la cohérence des connaissances représentées, leur qualification et leur enrichissement par raisonnement (Paulheim, 2017): vérification des types attribués aux entités topographiques, détection de faits contradictoires, déduction automatique de valeurs de propriétés, etc. Une attention particulière devra être apportée aux aspects de cohérence spatiale des données.

L'approche proposée sera testée sur deux cas d'applications :

Le premier, porté par le Service Hydrographique et Océanographique de la Marine (SHOM), s'inscrit dans la continuité de travaux récents sur la formalisation de connaissances sur l'environnement hydrographique marin et la navigation pour développer des services d'aide à la navigation à base de raisonnement : caractérisation automatique d'éléments du relief sous-marin à partir d'isobathes et de points de sondage (Yan et al., 2014) et (Yan et al., 2015), production automatique de cartes marines (Yan et al., 2017), description de trajectoires de navigation côtière (Laddada et Saux, 2017) ou encore génération automatique d'instructions nautiques (Haralambous et al., 2014) (Sauvage-Vincent et al., 2015) (Haralambous et al., 2017). Les approches proposées s'appuient sur une base de connaissances sur l'environnement hydrographique marin qui n'existe pas à ce jour. Les informations nécessaires à sa création sont disséminées dans différentes sources : ontologies construites manuellement, instructions nautiques en langage naturel, pilotes côtiers, référentiels de données géolocalisés, etc.

Le second, à définir selon la disponibilité des corpus, portera sur des textes règlementaires définissant des zonages sur le territoire français et tirera parti des référentiels de données géographiques produits par l'IGN pour le géoréférencement et la reconstruction des limites de ces zonages.

Direction de thèse, encadrement de thèse:

Directeur de thèse : Eric Saux (Irenav - Ecole Navale, eric.saux@ecole-navale.fr)
Co-directrice de thèse : Nathalie Abadie (LaSTIG - COGIT/Strudel, IGN, nathalie-f.abadie@ign.fr)
Encadrant : Eric Kergosien (Geriico - Université de Lille, eric.kergosien@univ-lille3.fr)

Contrat doctoral: Le contrat doctoral, d'une durée de trois ans, ouvre droit à une rémunération d'environ 1680 € brut (hors contribution aux frais de transports). Le contrat doctoral peut inclure pour l'ensemble de la durée de la thèse un service complémentaire d'enseignement, de diffusion de l'information scientifique et technique, de valorisation ou d'expertise.

Localisation: Equipe LaSTIG/Strudel, Institut national de l'information géographique et forestière (IGN), Saint-Mandé (métro 1, station Saint Mandé).
Début de la thèse: automne 2020

Profil recherché: Master 2 ou diplôme d'ingénieur en informatique : représentation de connaissances, Web sémantique, sciences de l'information géographique, extraction d'information à partir de textes.

Candidature : Envoyer par e-mail à l'ensemble des encadrants (1) votre CV, (2) une lettre de motivation adaptée au sujet proposé, (3) vos relevés de notes des deux dernières années d'études, (4) le cas échéant, l'avis du directeur de master (ou équivalent) et des lettres de recommandations.
Date limite de candidature: 15 mai 2020

Bibliographie:

Chen H., Vasardani M., Winter S., Tomko M. (2018). A graph database model for knowledge extracted from place descriptions. ISPRS International Journal of GeoInformation, vol. 7, no 6. Consulté sur http://www.mdpi.com/2220-9964/7/6/221

Gangemi, A. (2013) A comparison of knowledge extraction tools for the semantic web. In Extended Semantic Web Conference, pp. 351-366. Springer Berlin Heidelberg.

Haralambous, Y., Sauvage-Vincent, J. and Puentes, J., 2014. INAUT, a controlled language for the French coast pilot books instructions nautiques. In International Workshop on Controlled Natural Language (pp. 102-111). Springer, Cham.

Haralambous, Y., Sauvage-Vincent, J. and Puentes, J., 2017. A hybrid (visual/natural) controlled language. Language Resources and Evaluation, 51(1), pp.93-129.

Keller, A., N. Abadie, B. Dumenieu, S. Baciocchi and E. Kergosien (2018) Vers la construction d'une base de connaissances sur la réorganisation territoriale française à la Révolution, Actes de l'atelier Excès (Sagéo 2018).

Kim, J., Vasardani, M., Winter S. (2016) From descriptions to depictions: A dynamic sketch map drawing strategy, Spatial Cognition & Computation, 16:1, 29-53, DOI: 10.1080/13875868.2015.1084509

Laddada, W. and Saux, É., 2017. Description formelle de trajectoire de navigation en environnement maritime côtier. Revue Internationale de Géomatique, 27(2), pp.179-202.

Lamotte L., Abadie N., Saux E., Kergosien E..(2020) Extraction de connaissances pour la description de l'environnement maritime côtier à partir de textes d'aide à la navigation. Actes de la Conférence Francophone en Extraction et Gestion de Connaissances (EGC) 2020, pp.341-348

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, 8(3), 489-508.

Sauvage-Vincent, J., Haralambous, Y. and Puentes, J., 2015. Sentence ordering in electronic navigational chart companion text generation. In Proceedings of the 15th European Workshop on Natural Language Generation (ENLG) (pp. 66-70).

SHOM, France. Interest of semantic Web standards for S-100 developments. Paper for consideration by S-100 WG. En ligne, consulté le 01/04/2018 : https://www.iho.int/mtg_docs/com_wg/S-100WG/S-100WG2/S100WG2-9.3B_SemanticWeb.pdf

Suchanek, F. (2014) Information extraction for ontology learning.Lehmann and Völker [2 6] (2014): 135-151.

Yan, J., Guilbert, E. and Saux, E., 2014. An ontology for the generalisation of the bathymetry on nautical charts. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2(2), p.1.

Yan, J., Guilbert, E. and Saux, E., 2015. An ontology of the submarine relief for analysis and representation on nautical charts. The Cartographic Journal, 52(1), pp.58-66.

Yan, J., Guilbert, E. and Saux, E., 2017. An ontology-driven multi-agent system for nautical chart generalization.Cartography and Geographic Information Science, 44(3), pp.201-215.