Construction d'une mémoire des sites pollués

En vue de constituer un dossier de demande de financement auprès des financeurs possibles (IGN, ADEME, CEREMA et école doctorale VTT de l'Université Paris-Est), nous recherchons des candidatures pour le sujet présenté ci-dessous. Ce sujet pourra être adapté pour tenir compte des compétences et des intérêts de recherche de la personne candidate, avant d'être soumis aux financeurs.

Mots-clés

traitement automatique des langues, extraction d'information, information géo-référencée, entité nommée, événement

Contexte

La pollution est une des préoccupations centrales des Français, et en particulier des citadins. Les sources de pollution sont localisables, tout comme les zones exposées à ces risques. Le mode de diffusion et sa durée, l'étendue de la zone contaminée, dépendent du type de pollution envisagée (déchets industriels, radioactifs, de marées noires, de dragages, de guerre ; pesticides ; amiante ; polychlorobiphényles (PCB) ; etc.).
De nombreux acteurs produisent et/ou diffusent des informations concernant ces pollutions sous la forme d'indicateurs qualitatifs ou quantitatifs, de diagrammes, de textes (textes réglementaires, rapports techniques, arrêtés, décisions, comptes rendus de débats, etc.), de bases de données géographiques et thématiques (par exemples les bases BASIAS et BASOL du BRGM), de cartes et d'atlas (par exemple *Atlas de la France toxique*), de vidéos, etc. Ces acteurs, locaux, nationaux, européens, sont divers : institutionnels (villes, communautés de communes, ministère chargé de l'environnement, inspection des installations classées pour l'environnement, etc.), instituts ou établissements publics (Institut national de l'environnement industriel et des risques, Institut de veille sanitaire, Bureau de recherches géologiques et minières, etc.), organisations non gouvernementales, journalistes, associations de professionnels de santé, de scientifiques, d'usagers, etc.
Les supports des informations varient selon le producteur : journaux officiels, presse officielle, presse d'information sous toutes ses formes, sites officiels des instituts et organisations, sites collaboratifs, blogs... Enfin, ces informations, par leur nature, leur contenu, leur production, leur diffusion, évoluent dans le temps selon les événements concernant ces pollutions potentielles ou avérées.

Dans cette thèse, nous proposons de construire une mémoire des sites qui permette de rendre compte de la mémorisation ou de l'oubli collectifs des événements et activités d'un site, en rapport avec un risque de pollution. Il s'agit de collecter et organiser les informations produites à différents moments et par différents acteurs depuis l'installation d'une source polluante afin de construire des chronologies parallèles ; l'une relative aux événements (acte administratif, installation d'activité potentiellement polluante, changement d'activité, etc.), l'autre relative aux commentaires, prises de position, points de vue des acteurs en réaction aux événements.

Ce travail s'inscrit dans un champ multidisciplinaire qui couvre la linguistique de corpus et s'associe à la géomatique ; il utilisera/concevra des outils et méthodes de traitement automatique des langues (TAL) et des ressources et des outils de structuration et d'interrogation d'informations spatiales (bases de données géographiques, SIG, outils d'analyse spatiale).

La problématique de cette thèse serait double : d'une part, à partir des sources retenues, identifier les sites et les risques de pollution les concernant, d'autre part construire la mémoire de ces sites i.e. les événements relatifs aux sites et/ou aux risques de pollution. Les événements sont à comprendre comme, par exemple, l'installation d'une industrie ou d'une activité polluante ; les événements administratifs concernant ce lieu : permis de construire, de démolir ; arrêté d'autorisation, de mise en demeure, de suspension d'activité, de mesures d'urgence ; études de danger ; etc. ; les articles de presse concernant ces sites ou leurs activités ou les réglementations en rapport avec ces activités.

Les éléments pertinents s'articulent autour des lieux, combinés à des informations thématiques sur la pollution, la chronologie, les acteurs impliqués, les événements saillants. Il s'agira de développer des outils d'analyse des textes axés sur la détection de ces informations et en tenant compte de la pluralité des points de vue correspondant à des sources d'information différentes.

Verrous à lever :

Dans le domaine du traitement automatique des langues, ce travail relève de l'extraction automatique d'éléments d'information constituant un événement dans ce contexte de prévention des risques de pollution, et donc plus précisément d'informations concernant des lieux, des risques de pollution, des acteurs et des « actes » associés (production d'un document administratif, publication d'une information journalistique, diffusion d'un rapport technique, etc).

La notion de lieu est à rapprocher de celle d'entité nommée spatiale (ENS) qui est bien définie dans la littérature et il existe des outils permettant d'extraire et localiser automatiquement ces ENS. Cependant, les lieux visés par les risques de pollution sont souvent désignés par des noms communs (*l'usine*, *le lac, la décharge, la chaussée*) que ces outils reconnaissent mal (Brando *et al.* 2016). D'autre part, il ne s'agit pas de reconnaître tous les lieux mentionnés dans les sources d'information mais ceux en relation avec la thématique traitée.

La définition de la granularité d'un lieu est une question cruciale dans ce travail : par exemple, il n'est pas pertinent d'identifier *Paris* comme un lieu quand l'objectif est de différencier la Seine et les canaux de Paris à cause de leur pollution par les PCB, des voieries parisiennes et leurs enrobés amiantés à cause de la pollution à l'amiante.

Des travaux ont porté sur la détection des dates et des durées, par ex. (Teissèdre 2012 ; Moriceau & Tannier, 2014), pour la détection d'événements (Battistelli *et al.* 2013 ; Nguynen *et al.* 2016 ; Moriceau *et al.* 2017), pour la désambiguïsation d'entités nommées (Nouvel *et al.* 2015 ; Moreno *et al.* 2017 ; Melo & Martins 2017). Un verrou sera de classer ces événements pour en construire un déroulement chronologique à partir de dates saillantes (Kessler *et al.* 2012).

Enfin, afin de construire une mémoire d'un site ou d'un risque de pollution, il sera nécessaire de pouvoir associer un texte administratif (de portée générale) à un site ou un risque de pollution spécifique (Allan 2002).

Profil attendu

Master 2 (M2) ou équivalent en **informatique** avec un intérêt avéré pour le traitement automatique des langues, ou en **traitement automatique des langues** avec des connaissances et une pratique en programmation informatique et/ou gestion des connaissances et/ou web sémantique. La thèse relève d'intérêts pluridisciplinaires : informatique, traitement automatique des langues, information géographique.

Contrat doctoral

Après l'obtention d'un financement, le contrat doctoral d'une durée de trois ans ouvrira droit à une rémunération d'environ 1 700 € brut (hors contribution aux frais de transports). Il 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.

Toute candidature doit inclure :

1. un CV ;
2. une lettre de motivation adaptée au sujet proposé ;
3. un relevé de notes des deux dernières années d'étude ;
4. l'avis du directeur de master (ou de la personne responsable du diplôme donnant l'équivalence du master) ;
5. la copie du dernier mémoire ou rapport de stage, rédigé en français ou en anglais ;
6. le cas échéant des lettres de recommandation.

Bibliographie sommaire

Association Robin des Bois (2016). *Atlas de la France toxique*, Arthaud

Allan J. (2002). Introduction to topic detection and tracking. In *Topic detection and tracking*, James Allan (Ed.). Kluwer Academic Publishers, Norwell, MA, USA 1-16.

Battistelli D. (2011). *Linguistique et recherche d'information : la problématique du temps*. Hermès Science : Lavoisier, Paris.

Battistelli D., Charnois T., Minel J-L., Teissèdre C. (2013). Detecting salient events in large corpora by a combination of NLP and data mining techniques", In: *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics* (Cicling'13), 24-30 mars 2013, Samos, Grèce

Brando C., Dominguès C., Capeyron M. (2016). Evaluation of NER systems for the recognition of place mentions in French thematic corpora, In: *Proceedings of the 10th Workshop on Geographic Information Retrieval* (GIR '16). ACM, New York, NY, USA, article 7, 10 pages DOI: 10.1145/3003464.3003471

Kessler R, Tannier X., Hagège C., Moriceau V., Bittar A. (2012). Extraction de dates saillantes pour la construction de chronologies thématiques, *TAL*. Volume 52 – n° 2/2012, pages 57 à 86

Melo, F., Martins, B. (2017). Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, *21*(1), 3-38

Moreno J., Besançon R., Beaumont R., D'Hondt E., Ligozat A.-L., Rosset S., Tannier X., Grau G. (2017). Combining Word and Entity Embeddings for Entity Linking. In: *Proceedings of th14th Extended Semantic Web Conference* (ESWC 2017). Portorož, Slovenia, May 2017

Moriceau V., Tannier X. (2014). French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014).

Nguyen K-H., Tannier X., Ferret O., Besançon R. (2016). A Dataset for Open Event Extraction in English. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016). Portorož (Slovenia), May 2016

Nouvel D., Ehrmann M., Rosset S. (2015). Évaluation de la reconnaissance des entités nommées. In *Les entités nommées pour le traitement automatique des langues*, 111‑19. Sciences cognitives. London, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord: Iste éditions

Teissèdre C. (2012). *Analyse sémantique automatique des adverbiaux de localisation temporelle : application à la recherche d'information et à l'acquisition de connaissances*. Thèse, Université de Nanterre.