Mégadonnées, données liées et fouille de données pour les réseaux d'assainissement

Thierry Bonnabaud La Bruyère[1], **Nanée Chahinian**[1], **Carole Delenne**[1], **Laurent Deruelle**[2], **Mustapha Derras**[2], **Francesca Frontini**[3], **Rachel Panckhurst**[3], **Mathieu Roche**[4], **Lucile Sautot**[4], **Maguelonne Teisseire**[4]

1. *HSM, Univ. Montpellier, CNRS, IRD, Montpellier, France*
   nanee.chahinian@ird.fr
2. *Berger Levrault, Montpellier, France*

3. *Praxiling UMR 5267, CNRS, Univ. Paul-Valéry Montpellier 3, France*
   francesca.frontini@univ-montp3.fr,rachel.panckhurst@univ-montp3.fr
4. *UMR 9000 TETIS, Cirad, Irstea, CNRS, AgroparisTech, Univ. Montpellier Maison de la Télédétection, Montpellier, France*

RÉSUMÉ. *Le projet "Mégadonnées, données liées et fouille de données pour les réseaux d'assainissement" (MeDo) a pour objectif de tirer profit des mégadonnées disponibles sur le web pour renseigner la géométrie et l'historique d'un réseau d'assainissement, en combinant différentes techniques de fouille de données et en multipliant les sources analysées. Par l'amélioration des connaissances sur le réseau d'assainissement, ce projet contribue à une meilleure gestion du patrimoine hydraulique existant et de la ressource en eau et permettra d'analyser les interactions entre les politiques de développement urbain et les enjeux liés à la gestion de l'eau.*

ABSTRACT. *The "Megadata, Linked Data and Data Mining for WasteWater Networks" (MeDo) project aims to use Web big data for learning about geometry and history of wastewater networks, by combining different data mining techniques and multiplying analysed sources. The improved knowledge will lead to enhanced management of hydraulic heritage and water resources and allow analysis of interactions between urban development policies and water management related challenges.*

MOTS-CLÉS : *Fouille de données, mégadonnées, TALN, réseaux d'assainissement*

KEYWORDS: *Data mining, Big Data, NLP, wastewater networks*

1. Introduction

Les réseaux d'assainissement font partie intégrante de l'architecture urbaine. Durant le siècle passé, il était de coutume pour chaque opérateur de poser, d'entretenir et d'archiver les données relatives à son réseau (Rogers *et al.*, 2012). Avec les différentes politiques de privatisation/affermage/régie publique, les données ont souvent changé de lieu de stockage et de propriétaire. Pour certaines, notamment les plus anciennes qui n'étaient pas numérisées, l'information est généralement perdue.
L'objectif du projet « MeDo » est de tirer profit des mégadonnées disponibles sur le web pour renseigner à la fois la géométrie du réseau et les données « annexes » pouvant servir aux gestionnaires. Un premier défi consiste à traduire des informations textuelles non structurées en données quantitatives et en connaissance structurée du réseau. Le processus proposé s'appuiera sur des méthodes automatiques d'extraction d'information (EI) afin d'identifier des informations spatio-temporelles (Zenasni *et al.*, 2016) et thématiques (Frontini *et al.*, 2012) à partir des données textuelles en prenant en considération les typologies textuelles parfois non standard rencontrées (Roche *et al.*, 2016). L'analyse des incertitudes liées à ces informations est le deuxième aspect innovant du projet. En effet, il est nécessaire de transformer une information approximative en connaissance incertaine.

Bien que des applications de fouille de données existent déjà dans plusieurs domaines (médical, financier, reconnaissance de la parole et synthèse vocale), leur utilisation dans le domaine de l'eau est moins répandue (Chahinian *et al.*, 2016). Une exploitation originale des mégadonnées disponibles sur le web suppose une approche multidisciplinaire regroupant la linguistique et le traitement automatique du langage naturel (TALN) appliqué au français, l'informatique et l'hydrologie. Une mise en correspondance suivant les trois dimensions spatiales, thématiques et temporelles des données disponibles permettra de compléter les données de la base de référence et de restituer la dynamique locale du réseau au regard des différents acteurs impliqués et des ressentis des usagers, dans un objectif de prise de décision adaptée.

2. Matériels et méthodes

La chaîne de traitement proposée comporte une première phase de collecte de documents via le web pour la constitution d'un corpus exploitable. Les documents sont ensuite convertis en format texte, afin d'être exploités pour l'extraction d'informations sémantiques et spatiales.

Les documents sont récupérés à partir d'une succession de requêtes Google avec filtrage des résultats, ceux-ci étant limités aux dix premiers. Le programme

marque une pause de 90 secondes entre chaque requête afin de ne pas être bloqué par le serveur. La requête est de la forme :

```
"VILLE" AND ("EXPRESSION1" AND "EXPRESSION2") -MOTCLE1 -MOTCLE2
 -site:SITE1 -site:SITE2
```

Les textes contenant les expressions EXPRESSION1 et EXPRESSION2 en lien avec VILLE sont extraits. Certains mots-clés (bricolage, devis, plomberie, etc.) et certains sites (youtube.com, pagesjaunes.fr, etc.) sont exclus afin de ne conserver que les résultats les plus pertinents. La conversion des documents en texte brut se fait avec PDFMiner ou HTML2Text (deux modules Python), selon le type du fichier récupéré.

2.1. Extraction d'informations sémantiques

Une ontologie spécifique au domaine de l'hydraulique des réseaux d'assainissement est établie à partir d'une sélection par les experts de termes présents dans quatre lexiques disponibles sur le web [1].

Afin d'établir un corpus de comparaison et d'entraînement (gold standard), les documents trouvés sur le web sont classés automatiquement par genre textuel (officiel, presse, scientifique et social).

2.2. Extraction d'informations spatiales

L'extraction des entités spatiales et temporelles s'appuie sur la chaîne développée dans le projet Cart'Eaux (Delenne *et al.*, 2017) enrichie selon les besoins des experts et la base de référence développée. La mise en correspondance permet d'identifier des descripteurs et de construire des trajectoires retraçant la dynamique du réseau par une analyse a posteriori du cycle d'évolution de celui-ci (Fig.1).

Dans la chaîne de traitement mise en place, l'outil brat [2] est utilisé pour annoter les documents pour la phase d'apprentissage. Cette phase permet de valider l'extraction des entités nommées effectuée à l'aide de spaCy [3] (bibliothèque Python) et celle des entités temporelles faite avec Heideltime [4] (programme Java).

---

1. http://www.sivalodet.fr/accueil/lexique/var/lang/FR/rub/2921.html, http://www.hevia.fr/assainissement-lexique, https://www.siaap.fr/glossaire/lexique/B/, https://www.eau-anjou.fr/raccourcis/glossaire
2. http://brat.nlplab.org/
3. https://spacy.io/
4. https://heideltime.ifi.uni-heidelberg.de/heideltime/

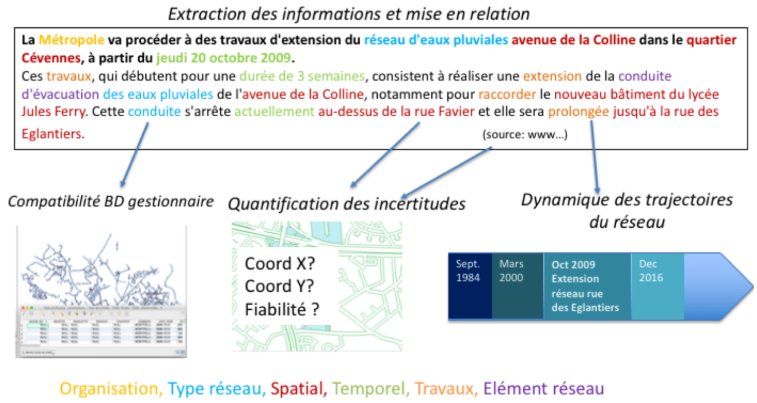

FIGURE *1. Exemple d'extraction des entités sémantiques et spatiales.*

3. Conclusions et Remerciements

Au-delà du web, le prototype développé dans le cadre du projet MeDo permet l'investigation de tout type d'archives numériques disponibles. La chaîne méthodologique mise en place est générique puisque son utilisation repose sur des mots-clés experts, communs à l'ensemble de la communauté des hydrologues, hydrauliciens, aménageurs et gestionnaires de l'eau. À l'avenir, nous pourrions approfondir cette recherche afin d'adapter le prototype à d'autres réseaux (naturels ou artificiels) ailleurs dans le monde, et ce au-delà de l'espace francophone.

Le projet MeDo, d'une durée de 24 mois, bénéficie du soutien de la **Région Occitanie-Pyrénées-Méditerranée** à travers le dispositif **« Recherche et Société(s) 2017 »**.

Bibliographie

Chahinian N., Piat-Marchand A., Bringay S., Teisseire M., Boulogne E., Deruelle L. *et al.* (2016). How can big data be used to reduce uncertainty in stormwater modelling? In *Spatial Accuracy 2016, 5-8 Juillet 2016.*

Delenne C., Chahinian N., Bailly J.-S., Bringay S., Commandre B., Chaumont M. *et al.* (2017). Cart'Eaux: an automatic mapping procedure for wastewater networks using machine learning and data mining. In *2017 AGU Fall Meeting.* New-Orleans, United States.

Frontini F., Aliprandi C., Bacciu C., Bartolini R., Marchetti A., Parenti E. *et al.* (2012). GLOSS, an Infrastructure for the Semantic Annotation and Mining of Documents in the Public Security Domain. In *Proceedings of the Workshop on Exploring and Exploiting Official Publications-EEOP2012 Istanbul May 2012.*

Roche M., Verine B., Lopez C., Panckhurst R. (2016). La néographie dans un grand corpus de SMS français : 88milSMS. In J. G. Palacios, G. D. Sterck, D. Linder, N. Maroto, M. S. Ibáñez, J. T. del Rey (Eds.), *La neología en las lenguas románicas Recursos, estrategias y nuevas orientaciones.*, p. 279-302. Peter Lang, Frankfurt.

Rogers C., Hao T., Costello S., Burrow M., Metje N., Chapman D. *et al.* (2012). Condition assessment of the buried utility service infrastructure – a proposal for integration. *Journal of Tunnelling and Underground Space Technology*, vol. 28, p. 331–344.

Zenasni S., Kergosien E., Roche M., Teisseire M. (2016). Extracting new spatial entities and relations from short messages. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, p. 189-196.