EGU21-2708, updated on 08 Mar 2021
https://doi.org/10.5194/egusphere-egu21-2708
EGU General Assembly 2021

WEIR-P: An Information Extraction Pipeline for the Wastewater Domain

Nanée Chahinian[1], Thierry Bonnabaud La Bruyère[1], Serge Conrad[1], Carole Delenne[1,2], Francesca Frontini[3,4], Marin Julien[1], Rachel Panckhurst[5], Mathieu Roche[6], Lucile Sautot[6], Laurent Deruelle[7], and Maguelonne Teisseire[6]

[1]HSM, Univ. Montpellier, CNRS, IRD, France.
[2]Inria Lemon, CRISAM - Inria Sophia Antipolis – Méditerranée, France.
[3]Instituto di Linguistica Computazionale "A. Zampolli" - CNR Pisa, Italy.
[4]CLARIN ERIC.
[5]Dipralang, UPVM, Montpellier, France.
[6]TETIS, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France.
[7]Berger Levrault, Montpellier, France.

Urbanization has been an increasing trend over the past century (UN, 2018) and city managers have had to constantly extend water access and sanitation services to new peripheral areas. Originally these networks were installed, operated, and repaired by their owners (Rogers et al. 2012). However, as concessions were increasingly granted to private companies and new tenders requested regularly by public authorities, archives were sometimes misplaced and event logs were lost. Thus, part of the networks' operational history was thought to be permanently erased. The advent of Web big data and text-mining techniques may offer the possibility of recovering some of this knowledge by crawling secondary information sources, i.e. documents available on the Web. Thus, insight might be gained on the wastewater collection scheme, the treatment processes, the network's geometry and events (accidents, shortages) which may have affected these facilities and amenities. The primary aim of the **"Megadata, Linked Data and Data Mining for Wastewater Networks" (MeDo) project** (http://webmedo.msem.univ-montp2.fr/?page_id=223&lang=en), is to develop resources for text mining and information extraction in the wastewater domain. We developed a specific Natural Language Processing (NLP) pipeline named **WEIR-P (WastewatEr InfoRmation extraction Platform)** which allows users to retrieve relevant documents for a given network, process them to extract potentially new information, assess this information also by using an interactive visualization and add it to a pre-existing knowledge base. The system identifies the entities and relations to be extracted from texts, pertaining network information, wastewater treatment, accidents and works, organizations, spatio-temporal information, measures and water quality. We present and evaluate the first version of the NLP system. The preliminary results obtained on the Montpellier corpus (1,557 HTML and PDF documents in French) are encouraging and show how a mix of Machine Learning approaches and rule-based techniques can be used to extract useful information and reconstruct the various phases of the extension of a given wastewater network. While the NLP and Information Extraction (IE) methods

used are state of the art, the novelty of our work lies in their adaptation to the domain, and in particular in the wastewater management conceptual model, which defines the relations between entities.