How can we build a corpus and an archive of the French web literature?

Christian Cote*[1]

[1]MARGE (MARGE) – Université Jean Moulin - Lyon III – MARGE Université Jean Moulin Lyon 3 18 rue Chevreul Bureau 510 - 5e étage 69362 Lyon Cedex 07 Tél. : 04 78 78 73 92, France

Abstract

We propose a workflow for data acquisition to build both a corpus and an archive of web literature. The corpus will be a tool for the exploration of the web literature containing an apparatus for information retrieval and navigation, then an indexation and metadata. The archive is a specialized archive that complete and enlarge the corpus but without files description.
The specificity of web literature is that we have no a priori criteria to determine what is literature. For that, we elaborate a methodology founded of the principle of mutual recognition among writers that found a workflow to explore the web and identify the URLs containing original literary production. This workflow is based on URLs exploration and strongly constrained web information retrieval. This methodology allows to characterize different types of recognition links and the detection of structured social networks. We use semantic rules to constrain search engines and classify issues in different modes of networks membership. The issues of this exploration have been recorded in XML schemas that will be used for the navigation in the corpus.

At a first time, we have crawled a corpus of literary production without any context (reviewers, editors, cultural institutions). The parametrization of HERITRIX(1) has allowing to restrict the number of jumps during the crawl and then the scope of the corpus. At a second time, we have enlarged the principles of identification by the use of a HYPHE method to detect each link associated to the identified URLs. In this way the issues of the first identification and the crawl have been evaluated. This second identification organizes more precisely the identified social networks and their boundaries. Thus, tenuous links to other sites appeared with the help of HYPHE(2), but also, some URLs identified during the first crawling but that have not been crawled due to restrictions. The second identification work made it possible to identify not only literary creation on the web, but literary life on the web, including the critical and social activities around literary creation. This difference precisely sheds light on the differences in the function and use of a corpus and an archive.

A corpus is a working tool that structures the data in such a way as to make explorations possible. We will explain the strategy adopted, centered on the users' needs and query terms to index files. We will present our difficulties and solutions for the automatic exploration of these WARC files for indexing purposes. Both collections used HERITRIX. However, the direct exploration capabilities of the collected data structured by WARC metadata made it necessary to duplicate the corpus with another collection, using the APIs of the platforms

*Speaker

and thus facilitating the automatic analysis. We will present how we use electronic lexical resources (LEXCONN and VERBNET in particular) to retrieve pre-defined textual types and ensure their extraction. This last work is in progress. We will present its first issues.

(1) https://github.com/internetarchive/heritrix3

(2) https://hyphe.medialab.sciences-po.fr/

Keywords: digital literature, specialized web corpus, archive building methodology