Hyphe: a curation-oriented approach to web crawling for the social sciences

Benjamin Ooghe Tabanou*[1]

[1]médialab Sciences Po – Sciences Po : médialab – 27, rue Saint Guillaume - 75337 Paris Cedex 07, France

Abstract

This communication introduces the RESAW's Archived Web studies audience to a curation-oriented web crawler tool named "Hyphe", which was developed with and for Social Sciences and Humanities scholars.
The World Wide Web's original design as a vast open documentary space built around the concept of hypertext made it a fantastic research field to study networks of actors of a specific field or controversy and analyze their connectivity and community structures along Bruno Latour's Actor Network Theory. Navicrawler, IssueCrawler, Hyphe... Over the past 15 years, a variety of web crawling tools, most often free and open source, have been developed by or for social sciences academics across the world. They provide means to engage with the web as a research field or to teach students what the web is beyond Google or Facebook's interfaces.

Developed by Sciences Po médialab as an open source software (https://github.com/medialab/hyphe), Hyphe was designed to provide researchers and students with a research oriented crawler to build and enrich corpora of websites through a qualitative fieldwork methodology. It provides a method and a tool to build a research corpus from web content (web pages and HTTP links) with an innovative "curation-oriented" step-by-step expansion approach meant to address two of the main social sciences problems when working with automatized web mining: how to build a theme focused corpus and how to delineate an actor's presence.

A step-by-step iterative process supports Hyphe users in dynamically curating and defining "web entities" in a way that is both granular and flexible by choosing single pages, a subdomain, a combination of websites, etc. The pages residing under these entities are then crawled, in order to extract the outgoing links and part of the textual contents. The most cited "web entities" can then be prospected manually in order to enrich the corpus before visualizing it in the form of a network and exporting it for cleaning and analysis in other tools such as Gephi.

We will complete this presentation with a demonstration of the latest version of médialab's "Hyphe-Browser", a complementary tool built as a fully functional web browser connected with Hyphe to let users benefit from automated web crawling as well as in situ web browsing and categorizing. Its friendly user interface allows a variety of publics to engage with web crawling, including non-experts like students, social science scholars or journalists.

*Speaker

Note: this communication presents how Hyphe allows to build corpora from the live web while another communication within the RESPADON panel will demonstrate how Hyphe was recently adapted to be able to also crawl Web Archives from archive.org and the national french library (BnF).

References:

- OOGHE-TABANOU, Benjamin, JACOMY, Mathieu, GIRARD, Paul & PLIQUE, Guillaume, "Hyperlink is not dead!". In Proceedings of the 2nd International Conference on Web Studies (WS.2 2018), Everardo Reyes, Mark Bernstein, Giancarlo Ruffo, and Imad Saleh (Eds.). ACM, New York, NY, USA, 12-18. DOI: https://doi.org/10.1145/3240431.3240434
- JACOMY, Mathieu, GIRARD, Paul, OOGHE-TABANOU, Benjamin, et al, "Hyphe, a curation-oriented approach to web crawling for the social sciences.", in International AAAI Conference on Web and Social Media. Association for the Advancement of Artificial Intelligence, 2016. DOI: https://doi.org/10.1609/icwsm.v10i1.14777

Keywords: webmining, crawling, hyperlinks, corpora