HyperText Corpus Initiative : how to help researchers sieving the web?

Paul Girard

HyperText Corpus Initiative : how to help researchers sieving the web?

Proposal for the "Using Web Archives" panel,
Out of the Box conference May 9, 2011.
by Paul Girard – médialab Sciences Po – paul.girard@sciences-po.fr

Since its foundation in May 2009, the médialab Sciences Po[1] works to foster the use of digital methods and tools in social sciences. With the help of existing tools and methods, we experienced the use of web mining techniques to extract data on collective phenomena. We also attended the symposiums organised by the two institutions responsible of web archiving in France: BnF[2] and INA[3] where we learnt about the difficulties posed to social scientists by the use of web archives. Actually our own experience in mining the live web wasn't easier. Such difficulties, we believe, can be explained by the lack of tools allowing scholars to build themselves the highly specialized corpora they need from the wide heterogeneity of the web. The web isn't a well-known document space for scholars or librarians. Its hyperlinked and heterogeneous nature requires to envision new ways of conceiving and building web corpora. And this notion of web corpus is a necessity for both live and archived web. If methods are not appropriate enough for analysing the live web, the problem will not be easier on an archive where the time dimension adds complexity.

In order to cope with this problem, we decided to launch an initiative joining the forces of actors coming from web archiving, web mining, social sciences and librarians communities. Our proposal is to set a coherent chain of tools to build, preserve and analyse web corpora. We aim at integrating existing pieces of software into a common methodological chain addressed at Social scientist and librarians needs. For this reason an important feature of this project is that it explicitly refuses to address huge corpora and targets. On the contrary, it focuses on topic-centric corpora in the order of a thousand nodes.

Thus the Hypertext Corpus Initiative[4] (HCI) has been founded in a kickoff workshop organised in October 2010 at the médialab Sciences Po gathering researchers, librarians and developers[5]. HCI working groups met several times in late 2010 to address the numerous issues our ambition has to face :

What is a website?
When one provides a researcher one of the existing web mapping tools, a question pops-up inevitably: what is a website? This question is far from being naïve. It comes from the difficulty encountered to map topic objects (actors, issues...) on URLs.
Indeed to analyse the web, the web pages have to be grouped in coherent sets, defined by a common portion of their URLs. The standard way to group pages is by domain

1 http://medialab.sciences-po.fr
2 http://*www.bnf.fr/documents/cp_memoire_web_sites_electoraux.pdf*
3 http://atelier-dlweb.fr/blog/
4 http://jiminy.medialab.sciences-po.fr/hci
5 from INA, BnF, Digital Methods Initiative Amsterdam, Density Design Lab Milano, Linkfluence, TIC migration research group, the Gephi consortium, Webatlas association, Sciences Po library and ISC PIF

name: http://www.domain.tld . But this way of grouping pages hides the heterogeneity of the web, as some websites are structured in subdomains or subfolders. Hence, grouping web pages by domain name is often irrelevant (blogs in a platform, profiles in a social network, papers in a press website...).
To address these concerns, we defined the concept of *web entity*. Our point is that web entities have to be defined by the user, based on what he considers a relevant set of web pages. The web entities are the basic aggregation level of web pages, the one on the top of which all other activities (like tagging, crawling, analysis, archiving…) will be based.

Control the corpus on a topic
The tools for building a web corpus currently at the disposal of social scientists (navicrawler[6] and issuecrawler[7]) impose a binary choice between manual and automatic approach. On one hand, the navicrawler is a brilliant tool to build a corpus out of a personal browsing experience. Even if this is the most accurate way of building a corpus, researchers were quickly discouraged by the amount of time this method requires. On the other hand, the online crawler issuecrawler (developed by the Digital Methods Initiative) is a perfect tool to expand a small list of URLs by automatically and remotely crawling them. However, the researchers feel they lost control on the corpus building process by discharging it to an automatic tool.
Finally neither manual nor automatic methods suit the need. It's like having to choose between tweezers and a caterpillar to find a needle in a haystack.
Thus we propose to develop a research driven crawling method, able to let the researcher control the crawl process. As we experienced, although automatic crawling is definitely necessary to harvest the outbounding links of a web entity, it's not suitable for extending the corpus. An automatic crawler should not be in charge of deciding which new web resources are to be included in or excluded from the corpus. This decision needs a qualitative analysis by the researcher. First to decide whether or not each new web resource fits the research intention but also to define the shape of the web entity to be associated to this resource.
For those reasons, we aim at developing a method of research driven crawling. A researcher should first define the granularity of his/her web entities, then decide to crawl them but with a depth-only crawling (i.e. crawling at null distance). While performing its crawling activities, the system should report new URLs back to the user asking him/her to create web entities from them. We believe that the researcher should analysethe results after each iteration of crawl to drive the automatic harvest by a qualitative analysis. A few quantitative indicators could be provided to guide the researcher in his/her decision making process. These are the very same indicators used for automatic filtering between iteration like co-link analysis or focus crawling, but used this time as a base for a manual decision.

Web corpus are not static maps
Due to the lack of integration of tools around a common notion of web corpus (at least in social sciences), the work results to static maps of websites. It's already a meaningful results but just the skin of what we could achieve.

6 http://webatlas.fr/wp/navicrawler/
7 https://www.issuecrawler.net/

First, the time dimension is missing. Here the collaboration with web archivists (BnF, INA and perhaps IIPC) is a key to foreseen the integration of archiving mechanism on the basis of the web corpus.
Secondly, corpus are outdated the minute they are built. Hosting a web corpus at the center of a series of tools let us imagine a way to frequently crawl the corpus not only to archive the selected web entities but also to detect new web resources as candidates to be included in the corpus. This process would then propose this new entry to the researcher who could decide to include it in or exclude it from the corpus.
Finally the notion of corpus will allow contents indexation and extraction to overtake the only topological analyses.

Explore to handle
To analyse web corpora we use an exploratory data analysis tool based on network visualisation (namely Gephi[8]). This is a perfect way to represent the nature and composition of the web. But why should we wait the end of the work to finally have exploration means at hands? We strongly believe that exploratory data analysis (EDA) could help researchers to handle the complexity of web corpora all along the process: seeing a map evolving during a crawl, using data visualisation to reflect quantitative indicators such as indegrees or terms frequencies, seeing clusters of web entities to understand the composition of the corpus… An important aim of our project is therefore to integrate such EDA tools in every step of the chain from the corpus building to the analysis. Those tools could be based on existing technologies, such as Gephi or processing[9], applied to the corpus of web entities created by researchers.

Following those principles, HCI develops a prototype called Hyphen. We hope this work will help fostering the use of web corpora in Social Sciences and thus the use of web archives by researchers. The first beta version is to be released in 2011.

8 http://www.gephi.org
9 http://www.processing.org