The Hypertext Corpus Initiative: methods and tools for Social Sciences to build corpus from the web

Paul Girard, Mathieu Jacomy, Audrey Baneyx, Tommaso Venturini

The Hypertext Corpus Initiative : methods and tools for Social Sciences to build corpus from the web

Paul Girard, digital manager, Sciences Po, médialab, Paris, France
paul.girard@sciences-po.fr
short bio : http://www.medialab.sciences-po.fr/en/team-en/paul-girard-en/

Mathieu Jacomy, web corpus specialist, Sciences Po, médialab, Paris, France
mathieu.jacomy@sciences-po.org
short bio : http://www.medialab.sciences-po.fr/en/team-en/mathieu-jacomy-en/

Audrey Baneyx, Data Center, Sciences Po Research, Paris, France
audrey.baneyx@sciences-po.fr
curiculum vitae in french : http://baneyx.net/?page_id=7

Tommaso Venturini, scientific coordination, Sciences Po, médialab, Paris, France
tommaso.venturini@sciences-po.fr
short bio : http://www.medialab.sciences-po.fr/en/team-en/tommaso-venturini-en/

Since its foundation in May 2009, Sciences Po's médialab has worked to enhance the use of digital methods and tools in Social Sciences. With the help of current tools and methods, we experienced the use of web mining techniques to extract and mine digital traces (hypertext links, spontaneous expression on blogs or social networks...) of collective phenomena. Our intention is to consider the web as a field to build new kind of corpora, and not as a research object in itself (web studies), neither as a media (innovative digital mediated surveys) nor as a medium (publishing or accessing structured digital data from the web).

This approach raised methodological and practical issues starting with the difficulty to build the highly accurate corpora needed by social scientists from the very complex document space that is the web : it has no size (too big, too dynamic), no clear boundaries because of its hyperlink structure and is composed of a wide heterogeneity of documents (technically, in usage, in time). How to qualitatively identify, select and collect web resources in such a quantitative context ? What does accuracy and representativity means in the moving matters of the web ? What are the tools which can equip the social scientists to build those new kind of corpora ?

Because we couldn't find a good enough answer to those questions by using the existing tools we decided to launch in October 2010 the Hypertext Corpus Initiative[a] gathering actors from web archiving, web mining, social sciences and librarians communities. HCI provides for social scientists a new set of methodology and tools, allowing them to mine more accurately digital traces of social phenomena from the web.

We will present in this paper the 4 mains methodological and technical issues discussed in HCI which lead us into developing a new set of tools :

(1) "what is a web corpus ?", introducing the concept of web entities to handle the complexity and heterogeneity of web resources;
(2) "how to build a web corpus ?", the methodological and technological issues regarding the quali-quantitative process of building a web corpus proposing to organize a research driven crawling for social sciences purposes;
(3) "how to analyze a web corpus ?", we would like to identified opportunities and limitations in using the web as a research field;
(4) "how to foster the use of web archives by social scientists ?", by applying web corpus principles to the archived web.

[a] http://jiminy.medialab.sciences-po.fr/hci