A network to develop the use of web archives : three outcomes of the ResPaDon project

Marie-Madeleine Geroudet*†1, Emmanuelle Bermes*‡2, Laurence Favier*§3, Eleonora Moiraghi¶4, Audrey Baneyx*5, Benjamin Ooghe Tabanou*6, Sara Aubry*7, Alexandre Faye*7, Amelia Laurenceau8, Antoine Henry3, Irène Bastard9, Ismail Timimi3, Joana Casenave3, Marie Cros10, Widad Mustafa3, Céline Ferjoux‖11, and Respadon Respadon**

[1]SCD, Université de Lille – Université de Lille, Sciences et Technolgies – France
[2]École Nationale des Chartes – Ecole Nationale des Chartes, École nationale des Chartes – France
[3]GERiiCO, Université de Lille – GERiiCO, université de Lille – France
[4]Direction des ressources et de línformation scientifique (Sciences Po) (DRIS) – Sciences Po – 27 rue Saint-Guillaume75337 Paris cedex 07 France, France
[5]Médialab (Sciences Po) (Médialab) – Sciences Po – 27 rue Saint-Guillaume - 75337 Paris Cedex 07, France
[6]Sciences Po médialab, DIME WEB – Fondation Nationale des Sciences Politiques [FNSP] – France
[7]BnF – Bibliothèque nationale de France – France
[8]Humathèque - Campus Condorcet – Campus Condorcet – France
[9]BnF – BNF – France
[10]SCD, Université de Lille – Université de Lille - Sciences et Technologies – France
[11]Respadon – BNF, Université de Lille, SCD – France

Abstract

This proposal targets a panel about the ResPaDon project, composed of 3 individual papers designed for a 15 mn presentation.
A highly transformative age requires new ways of doing research. In this context, web archives represent a huge opportunity for new types of research, offering possibilities for mining and analysis in many scientific disciplines. However, technical, legal, and methodological barriers can prevent researchers from using web archives in their research work. Among these barriers, the methodological cost of entry is the initial effort required from the researcher to get access to the collection and to understand the data available in the web archives.

The ResPaDon project (Network of partners for the analysis and exploration of digital data) aims in large part to reduce this methodological cost of entry, by associating academic and national libraries, researchers and librarians in a network of partners. The project is undertaken

*Speaker
†Corresponding author: marie-madeleine.geroudet@univ-lille.fr
‡Corresponding author: emmanuelle.bermes@chartes.psl.eu
§Corresponding author: laurence.favier@univ-lille.fr
¶Corresponding author: eleonora.moiraghi@sciencespo.fr
‖Corresponding author: celine.ferjoux@univ-lille.fr
**Corresponding author: WAS.conf@gmail.com

sciencesconf.org:resaw2023:434920

by the University of Lille and the National Library of France, in partnership with Sciences Po and the Campus Condorcet. It has been funded for 2 years by the GIS CollEx-Persée. It brings libraries and research teams together to think, experiment and share practices related to web archives. The main purpose is to bring the producers and the users of the web archive collection closer together, with the help and the mediation of academic libraries.

The goals of the project can be summarized as follows :

- To analyze the current and potential uses of web archives

- To experiment new ways of accessing and exploring web archives

- To issue recommendations about services, roles, skills and tools

This panel session presents three outcomes of the project. It shows how the methodology of the ResPaDon project leads to a better understanding of the place of web archives in the research process, to news ways of exploring web archives and to new ideas of services and training.

1) The use of web archives for scientific research

This paper presents the first results of a research project conducted on the use of web archives by French researchers. It has been undertaken according to two perspectives: the feedback from the BNF, which analyzes the evolution of research projects to which the digital legal deposit has been associated over the last 20 years, and an interview survey of researchers who have built their relevant corpus without specific assistance from library professionals.

The objective of this research is to identify and characterize the type of web source that interests scholars, the collection methods they use, their expectations with regard to the corpus they build, and the needs for processing tools necessary for scientific or teaching purposes that they express.

Based on a feedback drawn from 20 projects during the last 20 years, a typology of projects that have involved the BNF's digital legal deposit is proposed. The approach adopted consists in identifying the descriptors of the projects, then grouping those with common characteristics and thus defining "ideal-types". This method does not aim at uncovering mechanisms of particular projects or links, but at making a broader inventory in order to characterize distinctive elements.

An interview survey with scholars in political science, sociology, literature and history of the Web who have not been assisted by library professionals completes this typology of projects. It highlights several facets of long term research based essentially on Web materials: the simultaneous need for archives of the disappeared Web and those of the living Web, successive investigations to identify the source of content beyond the Web site, the need to collect materials from the entire Web ecosystem (including social networks), the involvement of private companies collecting the materials, the need to "make an archive", i.e. to build reference corpora that can be consulted and eventually updated over time. Completed by a teaching experience, putting in situation master students having to carry out research on various subjects, the difficulties to define the contours of the Web archive is emphasized.

The temporal dimension of the materials and of the tools for reading and collecting them, the shift from the collected "source" to the constituted archive bear epistemological as well as technical dimensions that make Web-based corpora scientific objects whose methodology has yet to be constructed. Scholars' feedback added to the experience of the BNF enable us to contribute into this endeavour.

2) Building and studying corpora from the Past Web using Hyphe

Since the end of 1990s, heritage and documentary institutions have built massive collections of digitized or born digital documents. Among these collections, Web archives have a unique place due to their complex documentary composition. Like the Web, they have opened up new opportunities for research and allowed a renewal of both methods of exploiting documentary corpora and models of interaction between information professionals and academic researchers.

One of the main focuses of the ResPaDon project is to provide the research community and information professionals with the possibility of appropriating methods and tools dedicated to the building, analysis and dissemination of web corpora. In this perspective, Sciences Po médialab and the National Library of France (BnF) organized, ran and evaluated an experiment based on the use of the Hyphe web crawler on Web archives.

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. It uses links between them in order to map web territories and enables the study of community structures. A step-by-step methodology supports Hyphe users in curating and defining "webentities" 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 webentities 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 in other tools such as Gephi.

As part of the ResPaDon project, Hyphe has been extended to work with the Past Web. The "Archives de l'internet", which is the name of BnF Web archives search application, and Hyphe are now able to work with one another. It allows the curation of both the live web and both BnF's and the Internet Archive's Web archives. Sciences Po and BnF also organized in April 2022 a one week event called a "datasprint", which brought together teams of researchers, engineers, designers, web archivists and collection specialists within the BnF DataLab, a new space and service dedicated to the development of Digital Humanities at the BnF. The purposes of this experiment were to determine if corpora building and curation software used on the live web can also operate on corpora from the "archived" web, and if logics of a comparative approach between the two are possible.

In this communication, we propose to draw up a first assessment of this experiment, to share the first questions raised by defining corpora from the Past Web and comparing them with the Present Web. We will also question the technical and epistemological potential offered by the application of methods and approaches from the Digital Humanities to Web archives collections.

3) Web archives : from heritage to science

Web archives, by nature plural and complex, are representative of the challenge we're facing when curating digital collections. Due to the global nature of the web, they are connected to all types of collections available in libraries and all research disciplines. Their technical characteristics make heuristic considerations essential both to their construction and to their understanding. Finally, they showcase the apparent contradictions of a legal framework that imposes access restrictions on content that was originally free. The Respadon project was born from these questions : its founding partners wished to lower the organizational, technical, legal and methodological barriers which limit the use of web archives as a source in French laboratories and research teams.

Among the main tasks of the project, a cycle of 8 workshops aimed at bringing together information professionals, researchers and other stakeholders (lawyers, training organizations, etc.) to imagine how web archives could be made more easily accessible throughout the country. By forcing us to translate our concerns from one profession to another, to identify our commonalities and our differences, the workshop series brought out questions so fundamental that we might tend to forget to bring them up.

First of all, are web archives really "archives"? If the term is well established among librarians, it remains confusing for potential users, faced with a methodological gap when approaching this reborn material. More than ever, it is crucial to open up this source and envision its use in relationship with other types of digital and analog material, and even more so with the live web, which remains an essential and complementary entry point.

Second, if access to web archives is too complex, we should just... make it simpler! Can we, as librarians, resist the temptation to explain, document, provide tools? Going back to simple concepts could prove life-saving. Entering into the web archives, for students and researchers, is a step-by-step process: they usually start by studying the content already available on the web, and then progress towards more complex tools such as text mining, link mapping...Providing clear and straightforward information on the library website, collaborating on seeds selections, building sandboxes or imagining "clandestine" ways of providing training are as many low-hanging fruits that we just need to seize.

Finally, those workshops led us to dream. Dream of the ideal training sessions for different audiences, dream of bringing together a vast and close-knit community, dream of lowering legal barriers, dream of web archives becoming an element of digital culture like any other, a recourse rather natural than mandatory for researchers in humanities. We hope to continue our work in the years to come to continue building this dream together.

References

- Ange Aniesa, Ariane Bouchard (2017) " Establishing an access network to the Internet archives : the French example", *IFLA WLIC 2017, Libraries, Solidarity, Society, Wroclaw, Pologne.* url : http://ifla-test.eprints-hosting.org/id/eprint/1655

- Emmanuelle Bermès (2019) "Digital uses and new relationships to museum data", *Quaderni*, Volume 98, pp. 73-86. url : https://doi.org/10.4000/quaderni.1455

- Sylvie Bonnel, Clément Oury (2014) "Selecting websites in an encyclopaedic national library: a shared collection policy for internet legal deposit at the BnF", *IFLA WLIC 2014 Libraries, Citizens, Societies : Confluence for Knowledge*, Lyon, France. url : http://library.ifla.org/id/eprint/998/7/
bonnel-en.pdf

- Gildas Illien, Pascal Sanz, Sophie Sepetjan, Peter Stirling (2011) "The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future", IFLA journal 38.1 (2012): 5-24. url : https://www.ifla.org/wp-content/uploads/2019/05/assets/hq/publications/ifla-journal/ifla-journal-38-1_2012.pdf#page=5

- Eleonora Moiraghi (2018) "The Corpus project and its potential audiences: A prospective study on the needs and expectations of future users", Bibliothèque nationale de France, Direction des Services et des Réseaux, 60 p., url: https://hal-bnf.archives-ouvertes.fr/hal-01739730

- Clément Oury, Karl-Rainer Blumenthal, Sébastien Peyrard (2016) "Digital Preservation Metadata Practice for Web Archives". In : *Digital Preservation Metadata for Practitioners.* Springer, Cham. p. 59-82.

- Stirling Peter (2017) "The Internet legal deposit in the Corpus project", *Web Corpora*, 24/05/2017, url : https://webcorpora.hypotheses.org/111

Keywords: Acces, experiment, services, skills, typology, survey, corpora, datasprint