A Network to develop the use of web archives: Three outcomes of the ResPaDon project

Sara Aubry, Audrey Baneyx, Emmanuelle Bermès, Laurence Favier, Alexandre Faye, Marie-Madeleine Géroudet, Benjamin Ooghe-Tabanou

Abstract: 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 work. The ResPaDon project (Network of partners for the analysis and exploration of digital data) aims to reduce the initial effort required from researchers to get access to web archive collections and understand them. It brings libraries and research teams together to think, experiment, and share practices, in order to analyze the current and potential uses of web archives, to experiment new ways of accessing and exploring corpora, and to issue recommendations about services, roles, skills, and tools..

Keywords: experimentation, services, datasprint, usage.

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. In France, the ResPaDon project (Network of partners for the analysis and exploration of digital data) works towards reducing this methodological cost of entry, by associating academic and national libraries, researchers, and librarians in a network of partners. This chapter presents three outcomes of the project, thus showing how our methodology leads to a better understanding of the place of web archives in the research process, new ways of exploring web archives, and new ideas of services, including teaching and learning activities.

Sara Aubry, Bibliothèque nationale de France (BnF), France, sara.aubry@bnf.fr, 0009-0009-6601-7546
Audrey Baneyx, Sciences Po Paris, France, audrey.baneyx@sciencespo.fr, 0000-0001-8500-9343
Emmanuelle Bermès, Ecole nationale des École nationale des chartes, France, emmanuelle.bermes@chartes.psl.eu, 0000-0002-7926-3027
Laurence Favier, Université de Lille, France, laurence.favier@univ-lille.fr, 0000-0002-5809-0788
Alexandre Faye, Bibliothèque nationale de France (BnF), France, alexandre.faye@bnf.fr, 0009-0007-8508-8492
Marie-Madeleine Géroudet, Université de Lille, France, marie-madeleine.geroudet@univ-lille.fr, 0000-0001-7816-0723
Benjamin Ooghe-Tabanou, Sciences Po Paris, France, benjamin.ooghe@sciencespo.fr, 0000-0001-7698-3507

1. About the ResPaDon project

The ResPaDon project, undertaken by the University of Lille and the National Library of France (BnF), received funding for two years from the GIS[1] CollEx-Persée. Initiated in early 2021, the ResPaDon partnership involved four libraries—the BnF, along with the three university libraries from Lille, Sciences Po, and Campus Condorcet—and two research laboratories: GERiiCOo (Groupe d'Études et de Recherche Interdisciplinaire en Information et COmmunication), an information science lab from the University of Lille, and the médialab at Sciences Po. The objective of bringing libraries and research teams together was to foster reflection, experimentation, and the sharing of practices related to web archives. The aim was to bridge the gap between the producers and users of the web archive collection, facilitated by academic libraries. The project reached its conclusion in April 2023, culminating with an international conference organized by the University of Lille.

From its inception, ResPaDon aimed to develop the analysis and exploration of collections as data (Padilla 2019), to equip researchers with computational tools for leveraging the digital collections of libraries, such as text and data mining capabilities. Following decades of digitization and digital legal deposit, the BnF and its partners successfully curated extensive digital collections that held significant research potential, especially in the humanities. Initiatives such as the CORPUS project (Moiraghi 2018; Stirling 2022) and the creation of the BnF DataLab, inaugurated in 2021 (Bermès 2019; Carlin 2021), gave rise to new services and collaborative approaches between librarians and researchers within the French national library. However, it became evident that their impact would remain limited unless they were distributed to a nationwide audience, far beyond the BnF premises, and in collaboration with other research libraries across France.

In our pursuit, amidst the massive amount of available materials, we decided to focus on web archives. Inherently plural and complex, web archives exemplify the challenges we encounter in curating digital collections. Due to the global nature of the web, these archives interconnect with all types of collections in libraries and across research disciplines. Their technical characteristics necessitate heuristic considerations, both in their construction and understanding. Finally, they bring to light the apparent contradictions of a legal framework that imposes access restrictions on content that was initially freely accessible (Stirling 2012), a situation that prompted prior initiatives at the BnF to create a national network of partners among legal deposit libraries in France (Aniesa and Bouchard 2017).

---

[1] Groupement d'Intérêt Scientifique, or Scientific Interest Group

The ResPaDon project emerged in response to these questions: its founding partners sought to dismantle the organizational, technical, legal, and methodological barriers hindering the use of web archives as a source in French laboratories and research teams.

To achieve this goal, the idea of a network was the cornerstone of ResPaDon. At the heart of the partnership, the collaboration between university libraries and a national library demonstrated the potential to build on a common professional culture, while maintaining different relationships with patrons and collections. One of the main tasks of the project was organizing a series of eight workshops to foster collaboration among information professionals and researchers, providing a platform to envision how web archives could be made more readily accessible throughout the country.

The project was organized into five work packages:

- WP1 Strategy: primarily focused on the workshop cycle,
- WP2 Understanding usage: dedicated to studying past and current uses of web archives and web materials,
- WP3 Capsule experimentation: an organizational and technical prototype of a web archive capsule at the University of Lille to provide secure access to the BnF's web archives,
- WP4 Live web and archived web: an experimental adaptation of a research web crawling tool to web archives, culminating in a week-long DataSprint co-organized by the Sciences Po médialab and the BnF DataLab,
- WP5 Planning, training and communication: a work package resulting in a series of events including the opening and concluding conferences.

The overarching goals of the project can be summarized as follows: to analyze the current and potential uses of web archives, to experiment with new ways of accessing and exploring web archives, and to provide recommendations about services, roles, skills, and tools.

2. The use of web archives for scientific research

Work package 2 of the ResPaDon project aimed at summarizing real and potential uses of web archives, with particular emphasis on researchers' practices. It was led by the GERiiCOlab at the University of Lille. The group approached its tasks from two key perspectives: firstly, feedback from the BnF, which analyzed the evolution of research projects associated with the digital legal deposit over the last 20 years, and secondly, an interview survey of researchers who built their relevant corpora without specific assistance from library professionals. This research aimed to identify and characterize the type of web sources that interest scholars, the collection methods they use, their expectations with regard to the corpora

they build, and their expressed requirements for processing tools essential to scientific or teaching contexts.

Based on feedback drawn from 20 projects spanning the last 20 years, a typology of projects involving the BnF's digital legal deposit was proposed. This study was built upon a previous survey from 2011 that comprised 15 interviews with researchers. This initial study had highlighted the necessity of web preservation, although the use of web archives was not significantly pronounced, and also revealed a growing interest in collaborative collection.

In 2022, the objective shifted to providing an extensive overview. The adopted approach involved identifying project descriptors then grouping them based on common characteristics, thereby defining 'ideal-types'. This method did not aim to uncover specific project mechanisms or links, but to create a broader inventory to characterize distinctive elements. Ultimately, 20 projects were analyzed using this approach and projected onto a timeline. We observed an increasing number of projects over time. Some projects were undertaken by individuals, others by organized research teams, but all shared a commonality: a multidisciplinary orientation towards humanities. In terms of the collections used, different categories of research emerged: some researchers were solely using the BnF's collection, others collaborated with the library on selection, and some mixed materials and resources from different origins.

Ultimately, five types of projects emerged, each detailed as ideal-types according to three criteria (a. public and objectives, b. service, organization, and duration, c. evaluation and improvement):

- **Punctual research:** carried out by an individual over a relatively short time span. The researcher freely accesses the web archive application to identify and evaluate content quality. The researcher works mostly autonomously.
- **Archiving and enrichment**: involves PhD students or scholars selected through a call process. It leads to long-term collaboration and results in valorization of the research through activities such as virtual guided tours ('parcours guidés').
- **Mining and exploitation**: teams of researchers engage in collective work, testing queries, and documenting datasets. Project selection includes a feasibility assessment.
- **Reference collection**: a laboratory aims to create a consistent collection related to their research topic. The produced corpus is indexed in full-text. The duration of the collaboration with the library can vary and there is often a need for increased visibility.
- **Production process**: a fully organized project with a team including IT skills. The project can lead to the production of a corpus and/or tools that may be standardized and distributed beyond the project stakeholders.

The GERiiCOlab complemented this typology through an interview survey with scholars in political science, sociology, literature, and history of the web who had not received assistance from library professionals.

An educational experiment was also conducted with a group of information science master students from the University of Lille. A teaching experience immersed them in research sessions particularly focusing on e-voting (with queries based on the French expression 'vote électronique' in the BnF web archive collection on 2002 elections) and femicides (using the term 'féminicide' in the BnF web archive news and media collection—collecte 'Actualité'). Students encountered challenges such as delineating the boundaries of web archives and distinguishing them from the live web. They also had difficulty understanding the typological difference between the web archives collected by the BnF and the 'archive' sections on certain websites, particularly those of news organizations. Choosing rigorous elements for analyzing web archives also posed problems due to the varying typology from one document to another, and their intersectionality with fields such as archives, documentation, and librarianship. Analyzing a website archived only on the surface (archived home page, but few—if any—pages available beyond the home page) raised issues in building web archive corpora. Moreover, students also faced challenges with the search interface, including searching by URL and understanding the outcomes of proposed tools.

Despite these difficulties, however, there were opportunities. Laurence Favier's focus on the BnF's News collection (collecte 'Actualité'), encompassing national and local press websites (pure player or cross media), aimed to identify researchers' needs. Capturing the full web environment of newspapers (including social networks) remains a challenge for web archive collections: from the production of new types of content (such as press blogs) to comments and, of course, links between other articles and new content. Web archives dedicated to newspaper websites actually compete with commercial news databases such as Europresse, for example. However, they offer researchers the ability to critique the live web and assess the current version of a website. It is a matter of authenticity: the web archive provides evidence when institutions decide what is outdated or or reliable on their websites.

Finally, the close interviews report touches upon the challenge of mixing research on the live web and on various propositions of archives/archiving. The ongoing evolution poses challenges to web archives processed within the legal deposit framework, yet it presents a real opportunity to share knowledge, tools, and methods for searching and building corpora. 'Web' and 'archives' actually interact in a variety of ways. The use of past and outdated content available online is not limited to institutional web archives such as those provided by the BnF, it may also include the online archives

section of a website, platforms managed by the community, or a private actor (including Internet Archive). Furthermore, researchers also conduct their own preservation actions, either in partnership with a heritage institution or by using online services to quickly save a page, such as Save Page Now or Archive Today.

To conclude, this segment of our project highlights several facets of long-term research primarily based on web materials. It emphasizes the simultaneous need for archives of the disappeared web and those of the live web, the successive investigations required to identify the source of content beyond the website, the need to collect materials from the entire web ecosystem (including social networks), the involvement of private companies in collecting the materials, the need to 'make an archive', i.e. to build reference corpora that can be consulted and updated over time. The difficulties when it comes to defining the contours of the web archive are accentuated, encompassing the temporal dimension of the materials and the tools for reading and collecting them. The shift from the collected 'source' to the constituted archive involves both epistemological and technical dimensions that turn web-based corpora into scientific objects whose methodology has yet to be constructed.

3. Building and studying corpora from the past web using Hyphe

One of the main focuses of the ResPaDon project was to provide the research community and information professionals with methods and tools designed for the creation, analysis, and dissemination of web corpora. In this perspective, the project's work package 4 involved a collaborative effort between the Sciences Po Library, médialab, and the National Library of France (BnF). Together, they organized, ran, and evaluated an experiment based on the use of the Hyphe web crawler on web archives.

Developed by Sciences Po médialab as open-source software, Hyphe[2] was designed to provide researchers and students with a research-oriented crawler for building and enriching website corpora (Ooghe-Tabanou 2018). The tool uses links between them to map web territories and allow for the study of community structures. Hyphe employs a step-by-step methodology that guides users in curating and defining 'webentities' in a granular and flexible manner by choosing single pages, subdomains, combinations of websites, and more. The pages beneath these entities are then crawled to extract outgoing links and part of the textual content. The webentities discovered through these links can then be manually selected and further explored to enrich the corpus in an iterative way. The corpus of webentities and interconnecting links can be viewed at all times in the form of a

---

[2] https://web.archive.org/web/20230120163329/https://github.com/medialab/hyphe

network and exported for cleaning and analysis in other tools such as Gephi. The outcome makes it possible to study online communities and social phenomena, identify and group web actors, and explore the links between them.

As part of the ResPaDon project, Hyphe underwent extensions to work with the Past Web. The 'Archives de l'internet'—the BnF web archive search application—and Hyphe are now compatible. Hyphe's code was also modified to enable its operation on the Internet Archive's Wayback Machine as well to facilitate the completion and/or comparison of past web corpora. The following developments were made:

- ability to crawl and explore a web archive using archival URLs or permalinks[3]
- ability to target a specific date and crawl documents within a defined time period, making sure not to drift over time[4] and target a specific date and time period around this date;
- support both the opened archival URL mode used by the Wayback Machine and the closed proxy mode used by the BnF's Archives de l'internet
- support the option of either building corpora from web archives only, or combining data from both the live web and the web archives within the same instance, even the same corpus. This flexibility enables the curation of both the live web and archived web, through the BnF's and the Internet Archive's web archives.

In April 2022, Sciences Po and the BnF co-organized a one-week event known as a 'DataSprint' (Venturini 2018), convening teams of researchers, engineers, designers, web archivists, and collection specialists within the BnF DataLab, a dedicated space and service for the development of Digital Humanities at the BnF. The purpose of a DataSprint is to gather complementary skills and expertise from a variety of disciplines (political and social sciences, web archiving, engineering, design, digital methods, etc.) and devote several full days to working collaboratively in small groups. The timeframe of a full week allowed participants to delve into data to explore hypotheses on specific research questions and experiment with different methodological approaches to gain preliminary insights. Through this experiment, participants aimed to assess whether the corpora building and curation software used on the live web could also operate on corpora from the archived web. Additionally, they sought to determine the feasibility of a comparative approach between the two.

---

[3] Such as http://archivesinternet.bnf.fr/20170401090112/http://www.lours.org/
[4] Time-drifting happens when a user (or, in that case, the Hyphe crawler) navigating through a web archive follows a hyperlink towards a resource that has been captured at a different date. If the target of the link is not present in the archive for the same capture date, the Wayback Machine will try to replace it with a capture from another date, thus taking the user to a period of time possibly far from their initial request.
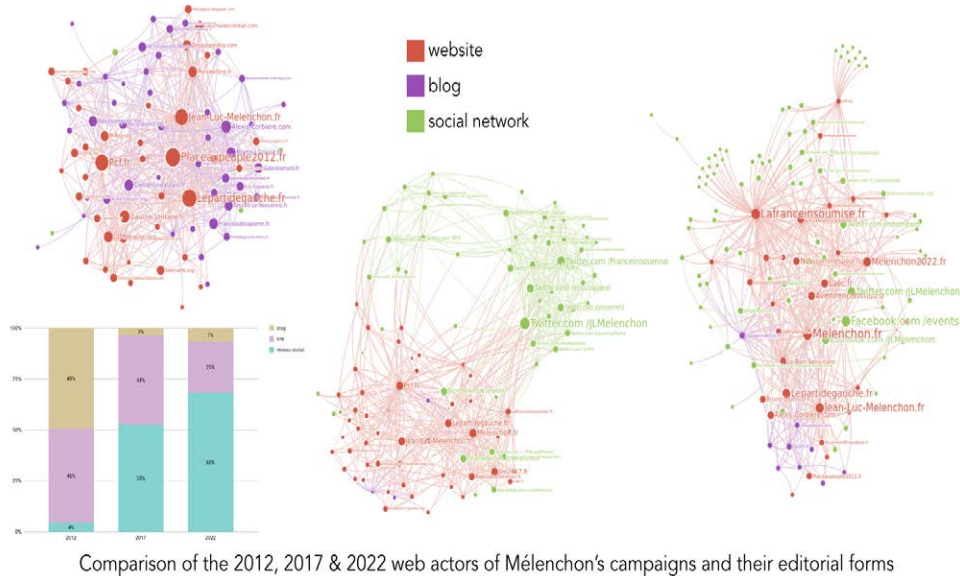
From April 4 to 8, 2022, twenty-five people worked within the BnF's DataLab reading rooms to explore and compare resources on the live web and web archives using Hyphe across four themes: political campaigns, COVID-19, theater arts critique, and genomics legal policies. The event resulted in the creation of a website[5] containing the DataSprint results with descriptions of the methodological tracks pursued, analyses conducted, and data visualizations generated.

The group working on the "Evolution of online political campaign structures" studied the structural transformations of online communities supporting candidates in the French presidential elections of 2012, 2017, and 2022, with a specific emphasis on Jean-Luc Mélenchon's campaigns. To facilitate a comparative approach, the group first defined the appropriate time windows based on archive availability. Coherent crawls were then initiated from persistent sources (Wikipedia pages), and common rules were established to select actors consistently across each snapshot. Finally, they devised a classification system to tag actors based on their editorial form and nature. With the corpora for the three consecutive presidential elections built, the challenge shifted to how to compare the compositions and structures of graphs. Digital methods were explored to visualize differences between two graphs, enabling the identification of disappeared actors, newcomers, and resilient ones. This allowed the group to compare the 2012, 2017, and 2022 web actors of Mélenchon's campaigns and their editorial forms, highlighting the subnetworks of websites involved in both the 2017 and 2022 campaigns and those exclusive to one of them. The compared graphs underscore the disappearance of blogs between 2012 and 2017, with the rise of social networks, initially forming as a separate network in 2017 and becoming deeply intertwined with campaign websites in 2022.

---

[5] https://web.archive.org/web/20231014053626/https://respadon.medialab.sciencespo.fr/

Comparison of the 2012, 2017 & 2022 web actors of Mélenchon's campaigns and their editorial forms

Another exploration during the DataSprint focused on the interactions between web officials and others on COVID-19. In this case, the comparative approach was based on three snapshots: 2020 (archives), 2021 (archives), and early 2022 (live web). The crawls were initiated from existing corpora built by collection curators and the group reused existing classifications to tag web entities by actor types (institutions, health experts, blogs, unions, NGOs, media, etc.). Their objective was to visualize the evolving position of an actor across the three graphs and the evolving proximity between actor types. They thus identified actors appearing, disappearing, or remaining over time, and explored how the positions of various actor types converge or diverge.

Methodological questions and future perspectives arising from this research include various uses for web archives in corpus building. These include longitudinal comparisons of time snapshots from the live web or archives, as well as the mixing of archives and live web to enhance corpora. Practical, technical, and systemic constraints have also surfaced such as the time-consuming nature of iterative corpus building and the complex reproduction process it entails. In this context, archive completeness and frequency is often an issue, and Hyphe has been considered a heuristic tool to evaluate it. Technical challenges include the permanent restructuring of the web (disappeared, repurposed, and redirected websites to new domain names), the resilience of links across time, and the complexities of modern full-JavaScript-based websites that are not yet crawlable with Hyphe. Building on these findings, Sciences Po médialab intends to pursue this work by connecting Hyphe to more existing web archives. Ongoing

collaboration with the INA (Institut national de l'audiovisuel) in France is already underway, and discussions with the national archives of other countries, such as Arquivo.pt in Portugal, have commenced following their expressed interest after the conference presentation.

4. Project results: main conclusions and recommendations

Circling back to the series of eight workshops organized within ResPaDon's work package 1, we can see how these ongoing research experiments on the use of web archives have fuelled a broader reflection on designing services tailored to academic needs. The workshops convened information professionals, researchers, and other stakeholders (engineers, lawyers, training organizations, etc.) to conceive how web archives could be made more accessible and usable for researchers. Each workshop session was organized around feedback presentations and discussions, covering eight topics throughout the series: access, usage, services, legal challenges, training, methods, the role of local and national players, and the creation and documentation of corpora. The outcome was a set of principles and recommendations regarding web archive access in higher education institutions, which were presented at the closing event in Lille in April 2023. The workshop findings were organized into a heuristic map and prioritized into potential actions based on five principles, as follows:

Principle 1. Given the unique nature of the web, the scientific study of its content necessitates the creation of an archive.
- Action 1. Support the definition and dissemination of methods for studying web sources to serve research.
- Action 2. Standardize the methodology for creating, documenting, and citing a web archive.
- Action 3. Integrate web sources into the development of digital literacy and culture for students, researchers, and professionals.

Principle 2. Web archives should serve as one of many sources for research.
- Action 4. Include web sources, archives, and the living web in the development of research practices and in opening up research processes and results.
- Action 5. Promote the exposure of metadata and the implementation of mechanisms for the discoverability of web archives.
- Action 6. Enable the discovery of web archives through browsing or exploring the living web.

Principle 3. Web archives should be usable autonomously for a variety of audiences.
- Action 7. Facilitate access to and reuse of web archives by changing current regulatory conditions.

- Action 8. Establish and sustain web archive access points in higher education and research establishments.
- Action 9. Facilitate the collective enrichment of an open-access sandbox for educational and research purposes.

Principle 4. A national network of researchers and information and library professionals is an essential catalyst for developing the use of web sources.

- Action 10. Unite actors interested in exploiting web sources and producing their archives around well-established nodes.
- Action 11. Develop various activities within these nodes to support access to and use of web sources, including awareness-raising and training, collaborative collections among partners, and remote access to web archives.
- Action 12. Implement co-leadership of the network by the nodes and national actors, involving shared documentation, regular meetings, and exchanges of practices.

Principle 5. The mediation of web sources by multiple actors requires the development of new skills.

- Action 13. Develop the skills of mediators within the nodes of the network.
- Action 14. Develop the skills of engineers supporting research projects based on web sources.
- Action 15. Develop collaborative collections by working with researchers and information and library professionals to encourage the acquisition of these new skills.

These recommendations represent just one facet of the project's deliverables. Notably, this chapter does not delve into the findings of work package 3, the experimentation capsule deployed in two locations within the university library in Lille. This real-life proof of concept, supporting the observations of researchers and students in work package 2, provided a comprehensive set of conclusions in terms of technical issues to secure access to web archives, organizational hurdles to overcome, as well as the need for documentation, training, communication tools, and support for these capsules to fulfill their role. Our recommendations also incorporate the results of this experiment as they seek to address the question: how can access to web archives be improved for researchers?

While ResPaDon's recommendations are meant to be realistic, we acknowledge that they are only drawing an ideal horizon, necessitating further consolidation in terms of resources. The partners have already convened to envision the next steps, including expanding the network to include new partners and addressing sustainability concerns for the service.

5. Conclusion

By compelling us to translate our concerns across the professions to identify our commonalities and our differences, the ResPaDon project raised questions so fundamental that they are often overlooked. Firstly, are web archives really 'archives'? While the term is well established among librarians, it remains confusing for potential users who encounter a methodological gap when engaging with this reborn material (Brügger 2018). It is now more critical than ever to open up and extend this source and envision its use in conjunction with other forms of digital and analog material, particularly with the live web, which remains a vital and complementary entry point. For students and researchers, entering the realm of web archives is a step-by-step process: starting with an exploration of the content readily available on the web and progressing towards more complex tools such as text mining and link mapping. Seizing the low-hanging fruits involves providing clear and straightforward information on the library website, collaborating on seed selections, building sandboxes, or devising discreet ways of providing training.

Ultimately, this project has inspired us to envision the ideal training sessions for different audiences, fostering a vast and close-knit community, dismantling legal barriers, and positioning web archives as an element of digital culture like any other, a resource that is natural rather than mandatory for researchers in humanities. As we look ahead, we hope to pursue our efforts in the years to come and continue building this dream together.

References

Aniesa, Ange, and Ariane Bouchard. 2017. "Constituer un réseau d'accès aux archives de l'internet : l'exemple français." https://library.ifla.org/id/eprint/1655/.

Bermès, Emmanuelle. 2019. "Quand le dépôt légal devient numérique : épistémologie d'un nouvel objet patrimonial." *Quaderni. Communication, technologies, pouvoir*, no 98: 73–86. https://doi.org/10.4000/quaderni.1455.

Bonnel, Sylvie, and Clément Oury. 2014. "Selecting websites in an encyclopaedic national library: a shared collection policy for internet legal deposit at the BnF." Paper presented at IFLA WLIC 2014 – Lyon – Libraries, Citizens, Societies: Confluence for Knowledge. Lyon, France, 2014. https://library.ifla.org/id/eprint/998/.

Brügger, Niels. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge (Mass.): MIT Press.

Brügger, Niels, and Ian Milligan. 2018. *The SAGE Handbook of Web History*. Newbury Park, California: Sage.

Carlin, Marie, and Arnaud Laborderie. 2021. "Le BnF DataLab, un service aux chercheurs en humanités numériques." *Humanités numériques*, no 4. https://doi.org/10.4000/revuehn.2684.

Gebeil, Sophie. 2021. *Website story: histoire, mémoires et archives du web*. Bry-sur-Marne: INA.

Gomes, Daniel, Elena Demidova, and Jane Winters. 2021. *The Past Web: Exploring Web Archives*. Switzerland: Springer Nature.

Harris, Kayla, Christina A. Beis, and Stephanie Shreffler. 2021. "Citizen Web Archivists: Applying Web Archiving as a Pedagogical Tool." *Journal of Electronic Resources Librarianship* 33, no 4: 262–72. https://doi.org/10.1080/1941126X.2021.1988463.

Milligan, Ian. 2020. "You shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure." *WARCNet papers*, Aarhus, Denmark. https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be__2_.pdf.

Moiraghi, Eleonora. 2018. "Le projet Corpus et ses publics potentiels." Rapport public. Bibliothèque nationale de France. https://hal-bnf.archives-ouvertes.fr/hal-01739730.

Musiani, Francesca, Camille Paloque-Bergès, Valérie Schafer, and Benjamin G. Thierry. 2019. *Qu'est-ce qu'une archive du web ?* Marseille: OpenEdition.

Ooghe-Tabanou, Benjamin, Mathieu Jacomy, Paul Girard, and Guillaume Plique. 2018.

"Hyperlink is not dead!" *Proceedings of the 2nd International Conference on Web Studies*, 12–18. https://doi.org/10.1145/3240431.3240434.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. 2019. "Always Already Computational: Collections as Data." https://zenodo.org/records/3152935.

Stirling, Peter. 2022. "Le dépôt légal de l'internet dans le projet CORPUS." Billet. *Web Corpora* (blog), 29 August 2022. https://webcorpora.hypotheses.org/111.

Stirling, Peter, Gildas Illien, Pascal Sanz, and Sophie Sepetjan. 2012. "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, no 1: 5–24. https://doi.org/10.1177/0340035211435323.

Venturini, Tommaso, Anders Munk, and Axel Meunier. 2018. "Data-sprinting: A public approach to digital research." In *Routledge Handbook of Interdisciplinary Research Methods*. Routledge. https://doi.org/10.4324/9781315714523-24.