Web archives
or
the Web as contemporary society's memory

Alexandre Chautemps,
Head of Dépôt Légal Numérique, Bibliothèque Nationale de France

Abstract:
We are living more and more on the Web. An increasing share of our discussions, production of ideas and artistic activities is happening on the Web and no longer on paper. There is, however, no guarantee that the Web's contents will survive. After a few years, it is no longer possible to find a webpage now available. For this reason, Web contents must be stored in reliable archives capable of reproducing operations as closely as possible to the Web. These archives will be an essential source of material for researchers who, in the coming years, will be studying our era. In fact, they have already started! After presenting a short history of the activities, questions and problems related to Web archives, attention is turned to the research programs on these archives. Lines of thought for the future…

A brief history of Web archives

Tim Berners-Lee invented the Web in 1989; and in 1993, CERN, his employer at the time, freely opened it for use. Barely ten years later, the Web became a very widely used means of communication. In 2004, 50% of the French had access to the Internet (BIGOT 2004, pp. 83ff). Individuals, society and institutions have already produced "natively digital" contents (*i.e.*, without a printed counterpart) of an impressive volume. Nonetheless, the idea of collecting documents from the Web to create lasting archives still seems peculiar. Only a few visionaries have devoted thought to this.[1]

Already in 1997, Brewster Kahle, the founder of the Internet Archive, wrote an article that, published in a major journal of popular science, stated the major principles of what would become the Web archives (KAHLE 1997). The Internet Archive and a few institutions (the Royal Library of Sweden as part of the KulturarW3 program and the National Library of Australia with Pandora) had made the first attempts to set up such archives in 1996. In France, the Bibliothèque Nationale de France (BnF) was, in 1999, experimentally collecting documents from the Web, but its first wide-ranging effort to do so was made during the election campaign in 2002 (ABITEBOUL *et al*. 2002). Following up on experiments done by the BnF and the Ministry of Culture, the French parliament decided in August 2006 to include in a bill of law a provision on the obligation of a legal deposit of e-documents (STIRLING *et al*. 2011). According to the Heritage Code, this depository

[1]This article has been translated from French by Noal Mellott (Omaha Beach, France). The translation into English has, with the editor's approval, completed a few bibliographical references. All websites were consulted in December 2020.

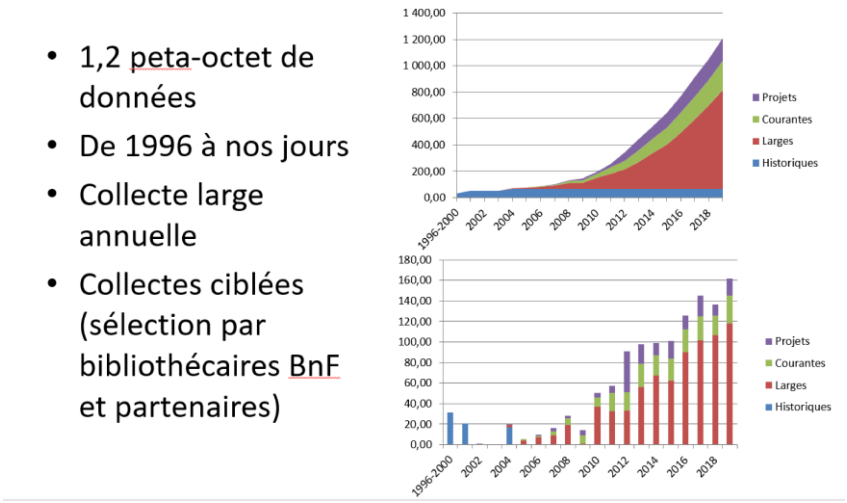
concerned "*the signs, signals, writings, images, sounds or messages of any sort that are communicated to the public by electronic means*" (Code du Patrimoine, Art. L31-2). This assignment was entrusted to three institutions: the Institut National de l'Audiovisuel (INA) for television and radio programs broadcast over the Internet, the Centre National de la Cinématographie (CNC) for films shown in theaters, and the BnF for everything else on the French Web.

It is worth emphasizing that the Web archive was assigned the status of a legal depository. This laid a legal basis for these activities with the implication of a cumulating legacy of e-documents (BERMÈS 2020, pp. 68ff). The "natively digital" collections of e-documents fit in with the BnF's collections of printed documents, the oldest of which date back to the 16th century, in the legal depository for documents on paper.

Every year, the BnF harvests all domains identified as belonging to the French Internet: all those ending in *.fr* but also other domains when the producer is located in France regardless of the extension (*.com*, *.org*, *.net*) and, too, the high level domains related to French overseas territories as well as new extensions (*e.g.*, *.paris*, *.bzh*, *.immo*, *.tools*).[2] In addition to this wide-reaching harvesting of e-documents, more frequent or deeper (and therefore more complete) harvesting campaigns target the sites that librarians at the BnF and in partner establishments have selected using documentary criteria. Some of these targeted campaigns are undertaken to keep electronic evidence of specific events (election campaigns, sporting events, attacks, epidemics, celebrations, etc.). Since 2010, the BnF uses its own equipment for this harvesting (LE FOLLIC *et al*. 2012). It oversees the storage and preservation of the digital legacy that is thus accumulating (DERROT *et al*. 2012).

By the end of 2019, the volume of the BnF's Internet archives amounted to approximately 1200 terabytes: nearly 35 billion files, the oldest of them dating back to 1996. This volume is growing day after day. The BnF's Internet Archives can be accessed on the premisses of the BnF and of twenty partner libraries.[3]

Figure 1: The Bnf's Archives of the Internet: Volume of collections at the start of 2020.
Source: © Bibliothèque Nationale de France



[2] For clarity about the names of domains and the extensions related to high-level domains, see the chapter "Noms de domaine et DNS" in Bortzmeyer (2018, pp. 89ff).

[3] https://www.bnf.fr/fr/archives-de-linternet. To locate the partner organizations, see https://bit.ly/2PjOpfF.

Using robots to harvest the Web has also turned up many e-books and e-journals, mostly in epub- or pdf-formats. harvesting does not target, however, the e-documents distributed by online retailers. Agreements have been signed with their publishers and distributors under the technically more appropriate procedure for legal deposits.[4]

Meanwhile, several other institutions — national libraries and archives, even foundations — are archiving the Web. Most of these institutions have joined the Internet International Preservation Consortium, which was set up in 2003, of which the BnF was a founding member (ILLIEN 2011). This consortium now brings together 57 institutions from 35 countries mainly in Europe, North America and, to a lesser degree, Asia.[5] Archiving the Web is still to begin in Africa, Latin America and India; this will be a major challenge in the coming decades.

Besides the mission assigned to "heritage institutions" of preserving Web contents for the future, the Web is also harvested to satisfy the immediate needs of research. A wide range of institutions do this, most of them not belonging to the IIPC consortium (mainly university libraries and research laboratories). The following section will focus on the uses of Web archives in research.

Use(s) of Web archives in research

As the millennium has advanced, natively digital documents, especially Web archives, have become legitimate sources for research in several human and social sciences. In fact, they have already become essential, indispensable sources for research work on the period since 2000.

Researchers' know-how is being reshaped by the digital humanities (MOUNIER 2010). Without giving up the aptitudes specific to their disciplines, historians, sociologists, anthropologists and linguists are becoming practitioners and sometimes specialists of natively digital data — identifying, selecting and manipulating them while taking an interest in the circumstances under which they have been produced and collected (MILLIGAN 2019). Web archives are thus a field of exploration of social groups and behaviors.

As a first example, I would like to mention Sophie Gebeil's project on the "memory" of North African immigration in France from 1999 to 2014 (GEBEIL 2015). Her work was centered not on events as such but on the "memorial process" they undergo. This approach, in line with the work of Pierre Nora (2008), seeks to reconstruct how French descendants of North Africans have drafted various narratives from the memory of their own individual or family histories. The BnF's Web archives were the principal source for this research. Thanks to the Web, relations were drawn between the discourses by: institutions (in particular, the Museum on the History of Immigration), the media, individuals, associations and/or activists. The growth of the Web has amplified the voice of these last categories. Sophie Gebeil's work has shed light on the creation of a heritage — the memory of immigrants, Harkis, neighborhoods, slums, and so forth — "memorial objects" that the conventional media have often overlooked or presented in stereotypes. This researcher choose to work on the archived Web, instead of the "live" Web, since her methodology required a stable corpus for counting occurrences, analyzing links between websites, observing the formation of networks within the community under study, etc. This made it necessary to have a corpus of texts that was "frozen" in time and protected from the volatility inherent in the live Web (GEBEIL 2017).[6]

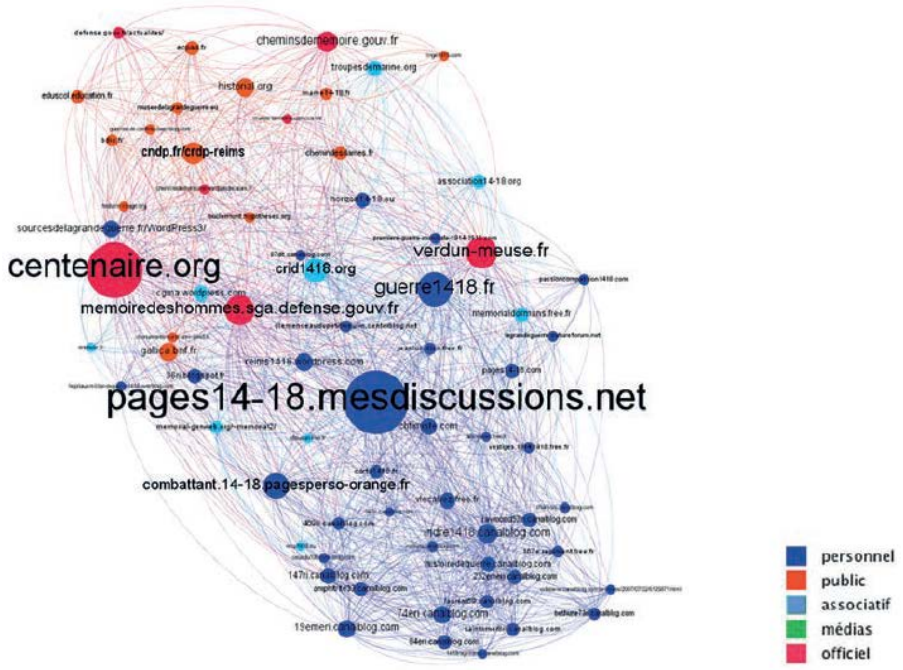
[4] Legal deposits of e-documents on line are being made experimentally under agreements with a publisher of e-books and a distributor of online music, respectively https://www.publie.net/ and https://idol.io/fr/.

[5] http://netpreserve.org

[6] See to https://www.bnf.fr/sites/default/files/2018-11/les_m%C3%A9moires_de_limmigration_maghrebine_parcours.pdf.

To take another example: *La Grande Guerre sur le Web*, a project that sought to see how the e-heritage related to World War I and placed on line by libraries has been reused on the Web, in particular during centennial commemorations (BEAUDOUIN *et al*. 2017 & 2018, CHEVALLIER 2017, BERMÈS 2020, p. 80).[7] The first step was to select the pertinent websites and describe them with a set of metadata defined for the project. The BnF then harvested the selected sites every semester to form a corpus for this research (total volume: 6.7 terabytes). The project's team used the software Gephi to map the relations woven between websites devoted to the history of the war. This set of websites was enlarged with those devoted to the commemoration. Some sites were institutional; others, made by amateurs or associations. Some focused on genealogy; others, on the history of regiments or on family narratives. Thanks to this mapping and statistics, the team brought to light not only two poles (institutional and "amateur" websites) that correspond to two quite distinct human networks but also a border zone where the two occasionally entered into communication (See Figure 2). During this project's second phase, Web Forum Pages 14-18, a frequent meeting place for exchanges, was thus targeted for harvesting.[8]

Figure 2: By Valérie Beaudouin, Télécom Paris, using the data from the BnF's harvesting campaign
Source: © Télécom Paris Tech



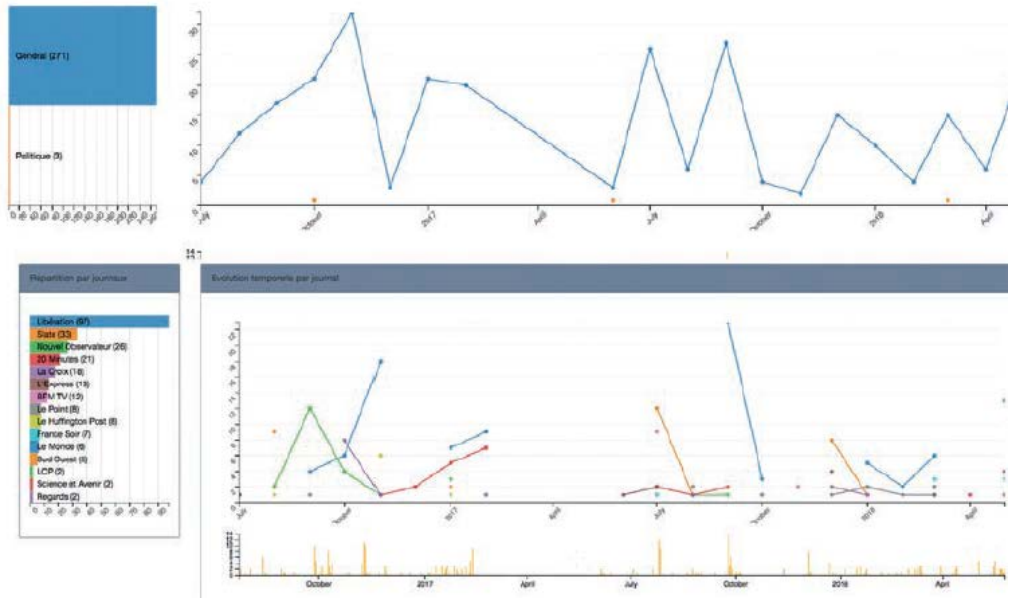
The Web also lends itself to studies of language change. Néonaute, a project from 2016 to 2018 headed by Emmanuel Cartier (Paris 13 University), sought to study the life cycles of three sets of words: 30,000 neologisms; French words recommended for usage in replacement of Anglicisms; and "feminized" words (AUBRY *et al*. 2018). The corpus came from the harvesting (Actualités) undertaken by the BnF between 2010 and 2017 (a volume of 13 terabytes). The pages accessible for free on French press websites were harvested daily. The corpus was fully indexed using Apache Solr and then processed to clean the webpages of anything (browsing links and buttons, headers, footers,

[7] A program from 2013 to 2018 headed by Valérie Beaudouin (Télécom Paris-Tech) and supported by Labex *Les passés dans le présent*. See https://www.bnf.fr/sites/default/files/2019-09/memoire_et_histoire_de_la_grande_guerre_sur_le_web_parcours_0.pdf.

[8] https://forum.pages14-18.com/

advertisements, etc.) that was not part of the text of articles. The team then undertook various morphosyntactic analyses, lemmatization and the extraction of text strings containing each occurrence of the words to be analyzed (but with enough of their textual environment for analyzing their function in the sentence). Changes in word usage were studied by taking account of the environment of a word's occurrence while distinguishing between the general press, news magazines, the technical press, etc.

Figure 3: Néonaute: A multidimensional, interactive visualization of occurrences.
Source: ©Paris 13 University



The Web is a prolific repository of scientific articles, many of them available in open access.[9] Given the many pdf-documents stored in Web archives, it is not easy to know which among them are scientific. A recent project, headed by Internet Archive, has used artificial intelligence (AI) to identify scientific articles in the mass of archived pdf-documents. The first step was to define, with the assistance of human librarians, a training corpus made up of the important sources of scientific publications in open access on the live Web. The machine learning algorithm was then trained to recognize which articles were scientific. Afterwards, it was used to process the archives to detect scientific articles among all the pdf-documents, evaluate the quality of the archived version (and archive documents if they were not already in the archives or if the archived copy was of poor quality), and signal the archived versions in Fatcat, an online catalog of scientific publications.[10] This project has convincingly demonstrated AI's capability for exploring the vast repositories of information in the Web archives, a task that, given the volume, people would be unable to do (PRAETZELLIS 2019).

[9] For definitions of the various sorts of open access, see https://espacechercheurs.enpc.fr/fr/open-access.

[10] https://fatcat.wiki

The Web archives are an indispensable source on the history of the Internet. This emerging discipline was initially told by the Web itself. Web archives enable us to reconstitute the succession of forms of technology (program languages, protocols, formats, and so forth) that have been used and to detect the major events in this history. The various periods and key moments in the Web's history have left tracks all over the archives: the passage from Web 1.0 to Web 2.0, the advent of blogs, the social networks, sharing platforms and recurrent trends in Web activism or Net art (BRÜGGER 2017, BRÜGGER *et al*. 2019, SCHAFER 2018a, 2018b & 2020).

A few prospects

As shown, the Web archives open a new field for investigation by researchers. This work involves a wide range of techniques for coping with an unprecedented volume of data. How to become familiar with these techniques and learn to use them? Several resources exist on the Web, of course.[11] Workshops ("datathons") are organized to introduce and train people to use these techniques for handling data, searching texts and visualizing data. It will be worthwhile to keep tab on the activities of Archives Unleashed, a group of Canadian historians who develop software and regularly organize datathons on historical Internet contents (MILLIGAN 2016).[12] Furthermore, libraries are setting up services for researchers who work on natively electronic data. A Web archives is one such service. The BnF is currently developing BnF Data Lab, which should be in operation by the end of 2020 (ELOI *et al*. 2019).

Although the use of such tools is indispensable, the wide diffusion of the digital humanities and the use of vast data collections in research will rely, above all, on human support.

References

ABITEBOUL S., COBÉNA G., MASANÈS J. & SÉDRAN G. (2002) "A first experience in archiving the French Web", *Proceedings of the Sixth European conference on Research and Advanced Technology for Digital Libraries* (ECDL), Rome, available via https://bit.ly/39JLpRv.

AUBRY S., CARTIER E. & STIRLING P. (2018) "Néonaute: Mining Web archives for linguistic analysis", slideshow presented at the IIPC Web Archiving Conference, Wellington, NZ., available via https://netpreserve.org/ga2018/wp-content/uploads/2018/11/IIPC_WAC2018-Sara_Aubry_Emmanuel_Cartier_Peter_Stirling-Neonaute-mining_web_archives_for_linguistic_analysis.pdf.

BEAUDOUIN V. & PEHLIVAN Z. (2017) "Cartographie de la Grande Guerre sur le Web: Rapport final de la phase 2 du projet *Le devenir en ligne du patrimoine numérisé: l'exemple de la Grande Guerre*", a report available at https://hal.archives-ouvertes.fr/hal-01425600.

BEAUDOUIN V., CHEVALLIER P. & MAUREL L. (2018) Editors of *Le Web français de la Grande Guerre: réseaux amateurs et institutionnels* (Nanterre: Presses Universitaires de Paris Nanterre)

BERMÈS E. (2020) *Le numérique en bibliothèque. Naissance d'un patrimoine: l'exemple de la Bibliothèque nationale de France (1997-2019)*, doctoral dissertation, École Nationale des Chartes, Paris, 107p., https://tel.archives-ouvertes.fr/tel-02475991.

BIGOT R. (2004), *La Diffusion des technologies de l'information dans la société française* (Paris: CRÉDOC), available at https://www.credoc.fr/publications/la-diffusion-des-technologies-de-linformation-dans-la-societe-francaise-2004.

BORTZMEYER S. (2018) *Cyberstructure: l'Internet, un espace politique* (Caen: C&F Éditions) referenced on https://catalogue.bnf.fr/ark:/12148/cb45637569v.

BRÜGGER N. (2017) Editor of *Web 25: Histories from the first 25 years of the World Wide Web* (New York: Peter Lang Publishing).

[11] One example among others: https://programminghistorian.org.

[12] https://archivesunleashed.org/aut/.

BRÜGGER N. & LAURSEN D. (2019) Editors of *The Historical Web and Digital Humanities: the Case of National Web Domains* (London: Routledge).

CHEVALLIER P. (2017) "Quand l'historien rencontre les Archives du Web", post on the blog Web Corpora, 30 January, available at https://bnf.hypotheses.org/1588.

DERROT S., FAUDUET L., OURY C. & PEYRARD S. (2012) "Preservation is knowledge: A community-driven preservation approach", *Proceedings of the Ninth International conference on the Preservation of Digital Objects (iPRES)*, Toronto, Canada, available via https://www.academia.edu/9753787/Preservation_Is_Knowledge_A_community_driven_preservation_approach.

ELOI C., MOIRAGHI E. & ROSE V. (2019) "Un espace pour les humanités numériques à la BnF", *Bulletin des bibliothèques de France (BBF*), 17, pp. 90-95, available at http://bbf.enssib.fr/consulter/bbf-2019-17-0090-009.

GEBEIL S. (2015) *La fabrique numérique des mémoires de l'immigration maghrébine sur le Web français (1999-2014)*, doctoral dissertation, 2 volumes, Aix-Marseille University.

GEBEIL S. (2017) "Quand l'historien rencontre les Archives du Web", post on the blog Web Corpora, 10 November, available at https://Webcorpora.hypotheses.org/380).

ILLIEN G. (2011) "Une histoire politique de l'archivage du Web", *Bulletin des bibliothèques de France*, 2, pp. 60-68, available at http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012.

KAHLE B. (1997) "Preserving the Internet", *Scientific American*, 276(3), pp. 82-83, avalable at https://www.jstor.org/stable/24993660?seq=1.

LE FOLLIC A., STIRLING P. & WENDLAND B. (2012) "Putting it all together: Creating a unified Web harvesting workflow at the Bibliothèque Nationale de France", paper presented at the IIPC workshop *How to fit in? Integrating a Web archiving program in your organization* held at the Bibliothèque Nationale de France in Paris from 26 to 30 November, available at https://hal-bnf.archives-ouvertes.fr/hal-00873759.

MILLIGAN I. (2016) "Lost in the infinite archive: The promise and pitfalls of Web archives", *International Journal of Humanities and Arts Computing*, 10(1), pp. 78-94, available at http://dx.doi.org/10.3366/ijhac.2016.0161.

MILLIGAN I. (2019) "Historians' archival research looks quite different in the digital age", *The Conversation*, 19 August, available at https://theconversation.com/historians-archival-research-looks-quite-different-in-the-digital-age-121096.

MOUNIER P. (2010) "Manifeste des *digital humanities*", *Journal des anthropologues* 122-123, pp. 447-452, available at https://journals.openedition.org/jda/352.

NORA P. (2008/1984) "Entre mémoire et histoire", article republished in P. NORA (editor) *Les Lieux de mémoire*, vol. I (Paris: Gallimard), pp. 23-43.

PRAETZELLIS M. (2019) "From open access to perpetual access: Archiving Web-published scholarship", slideshow presented at the IIPC Web Archiving Conference, Zagreb, available at https://netpreserve.org/ga2019/wp-content/uploads/2019/07/IIPCWAC2019-JEFFERSON_BAILEY__MARIA_PRAETZELLIS-From_open_access_to_perpetual_access-archiving_web-published_scholarship.pdf.

SCHAFER V. (2018a) *En construction: la fabrique française d'Internet et du Web dans les années 1990* (Bry-sur-Marne: INA).

SCHAFER V. (2018b) Editor of *Temps et temporalités du Web* (Nanterre: Presses Universitaires de Nanterre).

SCHAFER V. (2020) "Patrimoine, mémoires et histoire du Web dans les années 1990" post on the blog Web Corpora, 20 August, available at https://web90.hypotheses.org.

STIRLING P., ILLIEN G., SANZ P. & SEPETJAN S (2011) "La situation du dépôt légal de l'Internet en France. Retour sur cette nouvelle législation, sur sa mise en pratique depuis cinq ans, et perspectives pour le futur", paper presented at the *77th Congress of the International Federation of Library Associations (IFLA)*, Porto Rico, 29p., available via https://www.ifla.org/past-wlic/2011/193-stirling-fr.pdf.