De la Wayback Machine à la bibliothèque : les différentes saveurs des archives du Web …

Valérie SCHAFER

De la Wayback Machine à la bibliothèque : les différentes saveurs des archives du Web …
Filant la métaphore culinaire, cet article interroge le « changement de régime » qu'entraîne le recours aux archives du Web. Revenant à la fois sur les saveurs qu'elles dégagent, sur la manière de les servir et de s'en servir, il s'agit avant tout de penser la capacité de celles-ci à se marier à ce qui fait le sel de l'histoire et à composer des menus originaux. Ce parcours est une invitation à penser, à travers notamment des retours d'expériences sur une période qui va de 2011 à aujourd'hui, mais aussi d'autres références historiographiques, des enjeux tels que les techniques de fouille, d'analyse, de partage, mais aussi les limites qu'elles rencontrent.

Introduction

« Anyone who has lost track of time when using a computer knows the propensity to dream, the urge to make dreams come true and the tendency to miss lunch[1] ».

Comme le note le fondateur du *World Wide Web*, Tim Berners-Lee, il est aisé de perdre la notion du temps ou de rater l'heure du déjeuner, absorbé par une tâche informatique ou par la navigation dans la Toile, que ce soit dans le Web vivant ou celui du passé. Mais l'historien qui découvre les archives du Web ne reste pas longtemps sur sa faim… En effet s'ouvre alors à lui un univers singulièrement exotique, dépaysant, qui lui procure l'exaltation – et l'illusion, nous y reviendrons – de pouvoir à son tour naviguer dans le Web à la manière des premiers internautes des années 1990 pour les fonds les plus anciens.
Il est ainsi des moments dans une recherche que l'on n'oublie pas : la découverte des archives du Web en est un. Alors qu'en 2011 nous coordonnons avec Jérôme Bourdon un numéro spécial de la revue *Le temps des médias* dédié à la thématique « Histoire de l'Internet, Internet dans l'histoire[2] » et alors que je commence un projet consacré à la réception du Web en France dans les années 1990, rassemblant des matériaux traditionnels pour l'historien de l'innovation, tels des rapports étatiques, d'organisations, des archives de la presse généraliste et spécialisée, des archives

[1] BERNERS-LEE (Tim), « Interview par Kris Herbst », *Internet World*, juin 1994.
« Quiconque a perdu la notion du temps en utilisant un ordinateur connaît la propension à rêver, l'envie de réaliser ses rêves et la tendance à manquer le déjeuner » (notre traduction).
[2] BOURDON (Jérôme) et SCHAFER (Valérie) (dir.), dossier « Histoire de l'Internet, l'Internet dans l'histoire », *Le temps des Médias*, n°18, 2012.

audiovisuelles, ou encore une liste d'entretien oraux à réaliser, nous recevons une proposition de l'historien danois Niels Brügger. Il souhaite aborder les enjeux des archives du Web[3]. Je découvre alors la Wayback Machine[4] d'Internet Archive et la fondation états-unienne, créée en 1996 par Brewster Kahle.

Le premier contact avec la Wayback Machine est déconcertant, puisqu'il faut entrer une URL dans la barre de recherche avant de voir s'ouvrir les archives du site choisi. Sans surprise ni originalité c'est sur l'URL du CNRS, où je travaille alors, que s'arrête mon choix, et avec elle commence une exploration des archives du Web qui se poursuit jusqu'à aujourd'hui[5].

Revenir sur ce moment permet de mesurer le chemin parcouru par les archives du Web en l'espace de sept ans et les multiples changement en terme de méthodes d'archivage, d'accès, d'outils qu'elles ont connues[6].

En effet les archives du Web[7] ont différentes saveurs pour l'historien-ne, selon qu'il les consulte via la Wayback Machine en 2011 ou 2018. Au fil des années Internet Archive a introduit des fonctionnalités, permettant par exemple une recherche par mot clé sur les pages d'accueil des sites conservés, ou encore d'obtenir la date de collecte de chacun des éléments composant une page[8].

---

[3] BRÜGGER (Niels), « L'historiographie de sites Web : quelques enjeux fondamentaux », *Le Temps des Médias*, n°18, 2012, p. 159-169.

[4] Celle-ci permet de naviguer dans les archives du Web depuis 2001. http://web.archive.org

[5] MUSIANI (Francesca), PALOQUE-BERGÈS (Camille), SCHAFER (Valérie), THIERRY (Benjamin), *Qu'est-ce qu'une archive du Web ?*, OpenEditions Press, collection Encyclopédie numérique, à paraître, 2018.

[6] MERZEAU (Louise), « Vers un Web temporel », *Conférence du consortium international pour la préservation de l'internet* (IIPC), Paris, 19 mai 2014.
http://merzeau.net/vers-un-web-temporel/

[7] Si cet article n'a pas vocation à traiter dans le détail le statut de l'archive du Web, en termes archivistiques comme épistémologiques, mais se place plutôt du côté de l'usage de ces fonds par les chercheurs, nous soulignerons toutefois l'ambiguïté de la notion d'archive du Web, certes largement et internationalement utilisée, mais qui peut être questionnée. Ainsi, Bruno Bachimont dans *Patrimoine et numérique. Technique et politique de la mémoire* revient sur l'organisation des traces dans le cadre de l'archive, de la bibliothèque et du centre de documentation. Il rappelle que l'archive est conçue pour constituer « une preuve sur ce qui s'est passé » (p. 95). Mais « lorsque la constitution de l'ensemble documentaire obéit à une intentionnalité et un arbitraire lié non à la causalité de l'événement mais à la production des idées, on quitte le terrain de l'archive pour rejoindre celui de la bibliothèque » et donc celui des collections (p. 95). Les archives du Web s'apparentent ainsi plus à des collections liées au monde des bibliothèques et dans plusieurs pays au dépôt légal qui conserve des œuvres de l'esprit davantage que des traces d'activités. BACHIMONT (Bruno), *Patrimoine et numérique. Technique et politique de la mémoire*, Bry-sur-Marne, Ina, collection Médias et humanités, 2017, 246 pages, pp. 93-96.

[8] Voir le billet de GRAHAM (Mark), « Wayback Machine Playback… now with Timestamps! », *Internet Archive Blogs*, 5 octobre 2017.
https://blog.archive.org/2017/10/05/wayback-machine-playback-now-with-timestamps/

Mais les saveurs des archives du Web ne varient pas seulement en fonction des années et de l'ajout d'ingrédients, mais également en fonction du cadre de consultation des fonds, différent selon que l'on passe par la Wayback Machine, que l'on se rend à la Bibliothèque nationale de France (BnF) ou à l'Institut national de l'audiovisuel (Ina). Car les archives du Web créent ce semblant de paradoxe d'un retour à la bibliothèque, pour consulter ce patrimoine nativement numérique (en anglais *Born-Digital Heritage*). Disponible en ligne, à domicile ou sur son lieu de travail, dans le cas d'Internet Archive, il n'est accessible que dans les enceintes de la BnF et quelques bibliothèques en régions pour les archives du Web collectées par les institutions françaises dans le cadre du dépôt légal français (instauré en 2006[9]).
C'est un parcours à travers le goût des archives du Web, en livraison à domicile ou à la table d'une bibliothèque, que propose cette contribution, qui retrouve aussi la métaphore convoquée dans l'article « L'Ogre et la Toile[10] », par référence à la formule de Marc Bloch: « Le bon historien, lui, ressemble à l'ogre de la légende. Là où il flaire la chair humaine, il sait que là est son gibier[11] ». Mais après sept années de consommation régulière des archives du Web, c'est à une démarche plus intimiste et réflexive qu'invite ce papier, dans le sillage des ambitions de ce numéro.

De la friandise à la surabondance

Lorsque l'on découvre les archives du Web, on retrouve un peu des premiers réflexes maladroits qui étaient déjà perceptibles chez les minitélistes faisant leurs premiers pas en ligne: chercher son nom ou celui de son institution, entreprise, etc. Il s'agit aussi de découvrir sans but précis, de déambuler dans l'archive en se laissant guider d'hyperlien en hyperlien. Évidemment des liens sont brisés, des images ont disparu, mais on est plein d'indulgence face à ces archives, qui brusquement prennent le relai des captures d'écran patiemment rassemblées, par exemple au fil des dépouillements de la presse spécialisée (*Planète Internet, Internet professionnel*…), dépourvues d'interactivité et de profondeur.
Avec les archives du Web si l'illusion n'est pas totale, si l'historien découvre vite des sauts temporels étranges d'une page à l'autre, d'un hyperlien à l'autre, et une profondeur de collecte inégale, il a toutefois le sentiment de brusquement voir

---

9 COHEN (Évelyne) et VERLAINE (Julie), « Le dépôt légal de l'internet français à la Bibliothèque nationale de France », *Sociétés & Représentations,* vol. 1, no. 35, 2013, p. 209-218.

10 SCHAFER (Valérie) et THIERRY (Benjamin), « L'ogre et la Toile. Le rendez-vous de l'histoire et des archives du Web », *Socio*, nº 4, *Le tournant numérique... et après?*, coordonné par DIMINESCU (Dana) et WIEVORKA (Michel), 2015, p. 75-96.
http://journals.openedition.org/socio/1337

11 BLOCH (Marc), *Apologie pour l'histoire ou Métier d'historien*, Armand Colin, Paris, [1949], 1997, p. 4.

s'ouvrir un Web du passé plus vivant. Dès lors c'est le plaisir qui domine, celui de s'égarer dans la Wayback Machine, celui de chercher, à partir de différents annuaires tels celui de Yahoo! ou encore celui du Nic puis de l'Afnic (Association française pour le nommage Internet en coopération), des sites connus ou moins connus, de découvrir au fil d'une page un fameux « En construction » ou un gif encore plein de vitalité (pour satisfaire à ce plaisir gourmand du gif coloré, Internet Archive a d'ailleurs lancé en 2016 Gifcities[12]). On picore au grand buffet des archives du Web, entre sites institutionnels et pages plus personnelles qui auraient pu entrer dans la sélection dédiée au Web vernaculaire d'Olia Lialina[13]. Et on n'est pas loin d'éprouver cette nostalgie qu'a analysée Gustavo Gomez-Mejia à propos de Geocities et de Myspace ou au moins ce plaisir qu'il relève dans la « rusticophilie des écrans » : « Un charme désuet peut auréoler les écrans folklorisés de Geocities et Myspace dans la mesure où des souvenirs générationnels ressurgissent pour actualiser des perceptions nostalgiques[14] ».

On s'amuse aussi de la maladresse de certains sites, parfois vitrines de grandes institutions et pourtant bien moins réussis que ceux d'amateurs et qui peuvent conforter la formule : « Be small, look big ! ».

Et cette saveur de la découverte ne passe pas avec le temps. Nous l'avons collectivement retrouvée au sein du projet ANR Web90[15] en 2016 lors de la réalisation d'un parcours guidé dans les archives du Web des années 1990[16] à la BnF. Nous nous sommes régalés quand il nous a fallu sélectionner une centaine de sites, organisés thématiquement, par exemple autour de : État et services publics sur la Toile ; Le Web, modes d'emploi ; Les connecteurs ; Marchandisation et Net économie ; Les lieux de savoir ; Un terrain de créativité, etc.

Reste que ces archives du Web pour faire corpus doivent faire l'objet d'une sélection, appuyée sur une question de recherche mais aussi sur une démarche construite. Or, en 2011 la difficulté est réelle, alors que la Wayback Machine ne propose, comme nous l'avons souligné, que des recherches par URL, sans aucune visibilité sur la masse d'information conservée, sans inventaire des sites aspirés, sans information sur leur nature (problème que souligne également Jane Winters[17]), et que le plein texte

---

12 https://gifcities.org/

13 Voir notamment son site : http://art.teleportacia.org/observation/vernacular/

14 GOMEZ-MEJIA (Gustavo), « La fabrique de la désuétude. Regards diachroniques sur Geocities et Myspace », in SCHAFER (Valérie) (dir.), *Temps et Temporalités du Web*, Nanterre, Presses universitaires de Nanterre, 2018, p. 92

15 http://web90.hypotheses.org/

16 L'introduction de ce parcours est disponible à l'adresse :
http://www.bnf.fr/documents/web_annees_90_parcours.pdf

17 WINTERS (Jane), « Breaking in to the mainstream: demonstrating the value of internet (and web) histories », *Internet Histories*, 1:1-2, 2017, p. 173-179.

n'est pas implémenté dans la plupart des archives.
L'Ina propose déjà du plein texte, mais son périmètre d'archivage, dans le cadre du dépôt légal, est limité aux sites dont les contenus ont un lien avec l'audiovisuel. Ceux conservés par la BnF et par Internet Archive ne possèdent pas alors de recherche plein texte. Les choses ont bien changé depuis : la BnF a par exemple dans le cadre de projets internes (WebCorpus et Incunables du Web) implémenté la recherche plein texte en 2016 dans les fonds des années 1990, puis dans ceux des attentats de 2015. Elle propose également une liste des adresses URL des collectes ciblées du Web français[18] ainsi que des statistiques et métadonnées.
À un parcours lent, minutieux, appuyé sur le croisement des sources, recherchant les URLs pertinentes dans les guides de l'Internet, dans les rapports, dans les annuaires, dans la presse spécialisée, avec une curation humaine qui n'a rien à envier à celle de Yahoo! par exemple dans les années 1990, succède, en quelques années, avec le développement d'outils et la mise à disposition de métadonnées par la BnF et l'Ina une possibilité de déguster les archives du Web différemment, et même en quantité massive.
L'Ina par exemple a conservé des attentats de 2015 plus de 30 millions de tweets. Outre que le goût des archives du Web change forcément avec les archives de Twitter (voir dans ce numéro, la contribution de Frédéric Clavert sur le goût de l'API), qui ne reposent pas sur les mêmes recettes que les sites Web des années 1990, il n'est plus question pour le chercheur de picorage. Mais si la masse en elle-même peut vite mener à l'indigestion, c'est aussi évidemment le thème de la recherche, à savoir les attentats de 2015, que nous avons éprouvé en 2016 dans le cadre du projet interdisciplinaire ASAP (Archives Sauvegarde Attentats Paris[19]). Les réactions aux attentats et hommages sur Twitter défilent sur l'écran en quantité massive. S'affichent au fil des tweets des mots et des images que l'on aurait voulu éviter, à l'instar de la salle du Bataclan maculée de sang, des propos pro-Daech ou de ces messages de famille et d'amis à la recherche de proches, dont les visages surgissent au fil de l'archive et dont on connaît l'issue. Ces archives défilent *ad nauseam* au fil des archivages, présentations, affinages de corpus, etc.

Mise en bouche, menu complet ou plat de résistance ?

La question de la place des archives du Web dans le menu de recherche que concocte l'historien se pose aussi avec acuïté : les archives du Web sont-elles une mise en bouche, doivent-elles arriver après d'autres plats, comme cela nous était imposé en 2011, quand il fallait identifier via d'autres sources les URLs pertinentes, est-ce un

[18] http://www.bnf.fr/fr/collections_et_services/anx_pres/a.collectes_ciblees_arch_internet.html
[19] Voir le carnet de recherche dédié au projet : http://asap.hypotheses.org/

plat parmi d'autre ou peuvent-elles constituer le menu complet de la recherche ? Si à cette dernière proposition la réponse nous semble négative, tant il paraît difficile d'appuyer notre recherche historique sur une source exclusive, fut-elle pléthorique, et si nous avions pu défendre en 2015 dans « The Historian of the Web : Crawler, Browser or Lurker? » avec Francesca Musiani[20] le recours à une analyse qualitative, les fonds comme les outils ont fait quelque peu bouger nos lignes. Ce qui s'adaptait aux archives des années 1990 et à l'impossibilité en 2015 de rentrer dans les coulisses techniques de l'archivage a évolué grâce à la BnF en 2016, avec la fourniture de données statistiques générales et de nombreuses métadonnées. Pour autant cela n'a pas, dans ce cas de recherche précis, totalement modifié nos manières d'aborder la question : nous avons par exemple dû renoncer rapidement à des cartographies de liens. En effet l'enjeu des « missing data » est fondamental quand il s'agit de traiter de manière quantitative le Web archivé. Certains chercheurs ont ainsi essayé de mesurer la part du Web archivé par rapport au Web vivant qu'il tentait de conserver (cette part évidemment change dans le temps). Outre la recherche stimulante menée par Huuderman et al.[21] pour éclairer des méthodologies propres à saisir la représentativité du Web archivé, Hale et al. dans « Live versus archive: Comparing a web archive to a population of web pages », notent: « […] notre étude montre que les archives du web ne remplacent pas le besoin de collecter des données spécifiques de manière proactive sur des périodes spécifiques pour mener des études longitudinales. Face au degré d'incomplétude des archives du Web on peut même se demander dans quelle mesure elles peuvent être utilisées pour conduire des études longitudinales[22] » (notre traduction). Les archives des années 1990 de la BnF présentent une double limite tout particulière: elles ont été aspirées par Internet Archive de 1996 à 2000 de manière très lacunaire et sans que l'on puisse exactement repérer les critères de sélection ayant prévalu aux choix (certains sites nés avant même 1995 ne sont collectés qu'à partir de 2000) et la BnF au titre de sa mission de dépôt légal à partir de 2006 a récupéré ces archives d'Internet Archive sur une base de trois mois par an. Cette double sélection implique de limiter toute velléité d'obtenir des résultats, si ce n'est exhaustifs, même vraiment représentatifs du Web vivant de l'époque. Aussi effectuer une analyse quantitative sur un échantillon

---

20 MUSIANI (Francesca) et SCHAFER (Valérie), « The Historian of the Web : Crawler, Browser or Lurker? », in MILLIGAN (Ian) et WEBSTER (Peter), *Blog Web archives for historians*, 13 mars 2015. https://webarchivehistorians.org/2015/03/13/the-historian-of-the-web-crawler-browser-or-lurker/

21 HUURDEMAN (Hugo) et al.,« Lost but Not Forgotten: Finding Pages on the Unarchived Web », *International Journal on Digital Libraries*, 16(3), 2015, p. 247-265
https://pure.uva.nl/ws/files/2434657/163476_00799_015_0153_3.pdf

22 HALE (Scott), BLANK (Grant) et ALEXANDER (Victoria), « Live versus Archive: Comparing a Web Archive and to a Population of Webpages', in BRÜGGER (Niels) and SCHROEDER (Ralph) (eds.), *The Web as History,* London, UCL Press, p. 45-61. http://research.gold.ac.uk/20698/

lacunaire et opaque peut permettre d'obtenir des visualisions, de faire de l'analyse lexicale, mais les résultats obtenus ne parleront au mieux que du Web archivé, pas de l'ensemble du Web des années 1990 tel qu'il existait.
Inversement, l'archivage de Twitter au moment des attentats offre par son corpus pléthorique et documenté, la possibilité de parler de représentativité de ces archives au regard du flux en ligne. Les nombreuses fonctionnalités proposées par l'Ina offrent la possibilité d'obtenir très facilement des timelines, nuages de mots et graphiques (figure 1).

Figure 1 – Quelques résultats pour une recherche sur #prayforparis à l'Ina (timeline et diagrammes circulaires) © Ina

Il faut alors résister à l'appel du pré-cuisiné, alors que les outils peuvent très rapidement proposer des résultats qui risquent de ne pas avoir été correctement digérés par le chercheur.
Enfin « ingrédient, plat de résistance ou menu complet », la question reste également un enjeu majeur. Ayant commencé à travailler en amont de 2011 sur l'histoire du Web sans connaître les archives du Web j'avais commencé une histoire du Web sans archives du Web, grâce au recours à d'autres archives. La question est alors de savoir de quelle manière le recours aux archives du Web a changé l'écriture de cette histoire. Sans doute en permettant de porter par exemple davantage d'attention aux analyses visuelles, à l'architecture des sites, aux contenus, aux codes et langages dans la lignée des Code Studies, à la manière de naviguer de l'internaute de cette décennie, mais peut-être surtout en invitant à un travail plus collectif et interdisciplinaire, se devant d'emprunter aux sciences de l'information et de la communication ou encore aux humanités numériques, et en contact étroit avec les archivistes.
En effet, à la table de la bibliothèque peuvent ainsi s'inviter plusieurs disciplines, professions et expertises, en des projets interdisciplinaires, qui permettent de passer de la table solitaire du chercheur à une table d'hôtes prête au partage. La question du partage, et notamment celle du partage de corpus, reste aussi un enjeu majeur, qui dépasse les seules archives du Web pour toucher plus généralement au patrimoine nativement numérique[23].

[23] Par exemple sur les Newsgroups et forums en ligne : PALOQUE-BERGÈS (Camille), « Vers des lieux de mémoire réticulaires ? », *RESET* [En ligne], 6 | 2017, mis en ligne le 30 octobre 2016, consulté le 12 septembre 2018. http://journals.openedition.org/reset/ et PALOQUE-BERGÈS (Camille), *Qu'est-ce qu'un forum internet ? Une généalogie historique au prisme des cultures savantes numériques*, Nouvelle édition [en ligne], Marseille, OpenEdition Press, 2018.

Service à domicile ou table d'hôte : plusieurs recettes

L'archivage du Web est mené dans plusieurs institutions (voir la liste des initiatives d'archivage du Web sur Wikipedia[24] et la liste des membres de l'International Internet Preservation Consortium) selon des méthodes de capture, des interfaces et des présentations différentes, auxquelles s'ajoute aussi une diversité d'approches en termes de documentation et d'ajouts de métadonnées aux collections. Dès lors c'est aussi la relation du chercheur à ces archives qui va évoluer. Il faut s'adapter par exemple à des changements de terminaux, abandonner le Mac dont on est coutumier pour passer sur un PC dans une petite loge de la BnF, spécialement équipée pour accueillir l'équipe de recherche Web90 et celle d'ASAP en 2016 et fouiller grâce au plein texte ces fonds et les métadonnées.

Figure 2 – Recherche sur jesuischarlie via la plateforme Archives de l'Internet Labs mise en place par la BnF © BnF

Si les données et les outils mis à disposition compensent ces contraintes, elles ne sont pas toujours confortables, car ce sont aussi des changements de navigation, de manière de garder trace de son corpus et au-delà d'exploiter les résultats qui varient selon que l'on utilise ses propres outils ou ceux proposés par une institution tierce.
À première vue l'usage de la Wayback Machines pourrait sembler plus agréable, permettant notamment d'employer des outils choisis par le chercheur, de travailler où et quand il le souhaite, ou encore de faire des captures d'écran (ce qui ne garantit pas leur droit de diffusion ou de reproduction[25]). Mais l'historien-ne n'a pas forcément accès aux coulisses de l'archivage, à une vue d'ensemble, à des statistiques et métadonnées, ce qu'il va par contre trouver dans les institutions chargées du DL Web français. Par ailleurs les résultats diffèrent selon les fonds : après 2006, les collectes effectuées dans le cadre du dépôt légal par la BnF et l'Ina sont en effet bien plus fréquentes que celles de la fondation états-unienne pour certains contenus de la Toile française. C'est le cas pour les sites audiovisuels pour lesquels l'Ina offre de multiples captures quotidiennes, sans commune mesure avec celles d'Internet Archive, ou pour les archives des sites de presses, constituées pour la presse nationale

[24] Cf. la liste des initiatives d'archivage du Web sur Wikipedia (https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives) et la liste des membres de l'International Internet Preservation Consortium (http://netpreserve.org/about-us/members/)

[25] Cette fonctionnalité est par ailleurs offerte également pour les archives Web de l'Ina après validation humaine.

quotidiennement par la BnF et avec le choix de s'affranchir des robot.txt[26]. D'abord plongé dans les archives du Web en ligne, via son navigateur et son ordinateur, voici le chercheur qui revient à la table de la BnF, en un mouvement qui peut sembler paradoxal à l'heure où la numérisation peut sembler éloigner les chercheurs de la fréquentation des archives et bibliothèques, ou les voir y demeurer moins longtemps, occupés à photographier ou scanner plutôt qu'à traiter ces archives méticuleusement sur place. Une différence toutefois avec nos expériences antérieures est palpable: il s'y attablera parfois collectivement et avec les conservateurs. En effet désireux de saisir les besoins de chercheurs, pour adapter leurs collectes et outils, les archivistes et ingénieurs travaillent en interaction étroite avec les chercheurs et ce depuis par exemple les Ateliers du DL Web lancés par Claude Mussou et Louise Merzeau[27]. Le monde des bibliothèques et de l'archivage (archivistes mais aussi ingénieurs travaillant pour le DL Web) a été associé dans des projets aux équipes de recherche, que l'on pense par exemple à celui mené par la BnF avec Valérie Beaudouin autour des commémorations de la Grande Guerre[28] ou au projet ASAP précédemment mentionné mené avec la BnF et l'Ina. C'est un moyen aussi pour ceux qui oeuvrent à l'archivage du Web de saisir des attentes et de pouvoir adapter certains outils, comme le relève Zeynep Pehlivan (Ina):

> « Nos premiers développements d'outils se sont fondés d'abord sur une analyse des publications scientifiques et des méthodes convoquées dans ces travaux. Nous avons vu que la timeline par exemple était très utilisée, les agrégations et statistiques de base également. On voulait aussi permettre de filtrer le corpus, de créer des sous-corpus. Sur la version 1 de nos outils nous avons d'abord œuvré en cherchant à nous mettre à la place des chercheurs. Dans le premier lab on a eu un retour, des interactions, des demandes des chercheurs, par exemple celle de voir les emojis et leurs mentions, afin de pouvoir mener des analyses de sentiments, nous ne l'avions pas prévu. De même pour les nuages de mots. On a donc ajouté des fonctionnalités[29] ».

---

[26] Ce que ne faisait pas Internet Archive, qui par exemple n'archivait donc pas *Le Monde*. Toutefois en avril 2017 Internet Archive annonçait sur son blog avoir commencé à s'affranchir des contraintes imposées par les robot.txt sur certains sites. Pour plus d'explications voir le billet de Mark Graham : https://blog.archive.org/2017/04/17/robots-txt-meant-for-search-engines-dont-work-well-for-web-archives/

[27] Pour un témoignage sur ces ateliers, présenté lors du colloque anniversaire des 20 ans de l'archivage du Web, organisé conjointement par la BnF et l'Ina en 2016 à Paris, voir http://webcorpora.hypotheses.org/302
Le carnet de recherche des ateliers est par ailleurs consultable à : http://atelier-dlweb.fr/blog/

[28] BEAUDOUIN (Valérie) et PEHLIVAN (Zeynep), *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 »,* [Rapport de recherche] Bibliothèque nationale de France; Bibliothèque de documentation internationale contemporaine; Télécom ParisTech. 2017. <hal-01425600>

[29] Entretien avec Zeynep Pehlivan (Dépôt Légal du Web, Ina), 2017.

Conclusion

Dans le cadre de ce projet évoqué par Zeynep Pehlivan, celui autour de l'archivage de Twitter lié aux attentats qui frappent la France en 2015 et 2016, le sujet abordé est très contemporain. Les archives du Web doivent-elles se consommer fraîches ou au contraire se bonifient-elles en vieillissant ? Le chercheur qui mène une étude diachronique d'un site va devoir affronter des archives de qualité très différentes, du « gruyère » des années 1990 à celles plus complètes et roboratives de la période récente, dont les images, mais aussi le son, les vidéos, etc. sont mieux conservées au fil des années.

Une autre problématique est associée à cette question : plus l'écart va se creuser entre la collecte des archives du Web et leur exploitation, plus il y a de chance que le chercheur ne soit pas forcément au fait de tout le contexte matériel mais aussi systémique du Web vivant tel qu'il s'inscrit dans l'époque qu'il étudie. C'est donc aussi un Web en contexte[30] qu'il lui faudra pouvoir restituer, attentif aux pratiques, équipements, usages, lieux de connexion, etc. pour ne pas risquer la perte d'intelligibilité[31].

Demeure enfin, alors que l'évolution de l'archivage du Web, mais aussi sa place dans l'offre de formation des historiens sont loin d'être stabilisées, la question des risques d'entorses au régime non pas seulement de vérité[32], mais aussi éthique que l'historien doit affronter. Autant d'enjeux qui promettent encore quelques défis et invitent à une cuisine du monde, inventive et collaborative.

Valérie Schafer
Professeure d'histoire européenne contemporaine
C²DH, Université du Luxembourg
valerie.schafer@uni.lu

---

30 SCHAFER (Valérie) et THIERRY (Benjamin), « Web History in context », in BRÜGGER (Niels) and MILLIGAN (Ian), *The SAGE Handbook on Web History*, SAGE, à paraître 2018.

31 Ce risque est souligné notamment dans BACHIMONT (Bruno), *Op. Cit.*

32 RICOEUR (Paul), *Histoire et vérité*, Paris, Points Essais, 2001, 416 p.